National Library
of Canada

Canadian Theses Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale
du Canada

Services des thèses canadiennes

CANADIAN THESES

THÈSES CANADIENNES

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

NL 339 (r. 86/01)

Canada

National Library of Canada
Bibliothèque nationale du Canada

Ottawa, Canada
K1A 0N4

64

TC –

0-315-22930-6

CANADIAN THESES ON MICROFICHE SERVICE – SERVICE DES THÈSES CANADIENNES SUR MICROFICHE

## PERMISION TO MICROFILM – AUTORISATION DE MICROFILMER

• Please print or type – Écrire en lettres moulées ou dactylographier

### AUTHOR – AUTEUR

Full Name of Author – Nom complet de l'auteur
SHERRIE ELLEN SHAMMASS

Date of Birth – Date de naissance
JULY 21, 1957

Canadian Citizen – Citoyen canadien
☑ Yes / Oui ☐ No / Non

Country of Birth – Lieu de naissance
CANADA

Permanent Address – Résidence fixe
#209 - 7431 BLUNDELL ~~RICHMOND~~ ST.
RICHMOND, B.C.

### THESIS – THÈSE

Title of Thesis – Titre de la thèse
FORMANT TRANSITIONS, SPECTRAL SHAPE, AND VOWEL CONTEXT IN THE PERCEPTION OF VOICED STOPS

Degree for which thesis was presented / Grade pour lequel cette thèse fut présentée
PhD

Year this degree conferred / Année d'obtention de ce grade
1985

University – Université
UNIVERSITY OF ALBERTA

Name of Supervisor – Nom du directeur de thèse
DR. T. M. NEAREY

### AUTHORIZATION – AUTORISATION

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to **microfilm** this thesis and to **lend** or **sell copies of the film.**

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de **microfilmer** cette thèse et de **prêter** ou de **vendre des exemplaires du film.**

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

**ATTACH FORM TO THESIS – VEUILLEZ JOINDRE CE FORMULAIRE À LA THÈSE**

Signature
Sherrie Shammass

Date
Oct. 11, 1985

NL-91 (r. 84/03)

Canada

THE UNIVERSITY OF ALBERTA

# Formant Transitions, Spectral Shape, and Vowel Context in the Perception of Voiced Stops

by

© Sherrie E. Shammass

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF Doctor of Philosophy
IN
Speech Production and Perception

Department of Linguistics

EDMONTON, ALBERTA
Fall, 1985

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Sherrie E. Shammass

TITLE OF THESIS Formant Transitions, Spectral Shape, and Vowel Context in the Perception of Voiced Stops

DEGREE FOR WHICH THESIS WAS PRESENTED Doctor of Philosophy

YEAR THIS DEGREE GRANTED Fall, 1985

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) Sherrie Shammass

PERMANENT ADDRESS:

11A Hatserot Hadar
Kfar Saba, Israel
44359

DATED Oct. 7 1985

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Formant Transitions, Spectral Shape, and Vowel Context in the Perception of Voiced Stops submitted by Sherrie E. Shammass in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Speech Production and Perception.

..Terrance M. Nearey..........

Supervisor

...John T. Hogan...

............................

...Bruce L. Derwing...

Jamieson

............................

External Examiner

Date....7 Oct. 1985.....................

*To Mama, Papa, and Sasa*

# Abstract

This study investigates the perceptual importance of formant transitions, spectral shape and vowel context in the identification of place of articulation for initial voiced stop consonants. The formant onsets and steady-states were measured for /b/, /d/ and /g/ in eleven English vowel contexts for both male and female speakers. For each stop category, linear trends were evident when onsets were plotted vs. steady-states; linear regression lines were obtained. Classification scores based on these linear regression lines gave 73% correct classification. Perception studies using synthetically generated speech were done in which the formant onsets were systematically manipulated in five steady-state contexts. Of these five contexts, three were heard as unambiguous vowels /o/, /U/, or /ɛ/, and two were ambiguous: [o-U] and [U-ɛ]. Consonant categorizations were different depending upon the F2 steady-state context. An analysis of the roles of formant transitions and onset spectral shapes indicated that formant-based information was important for predicting subjects' /d/ responses, while spectral shape seemed to better predict their /g/ responses. Models based on either formant or spectral shape information made better predictions for stimuli with higher F2 steady-states. An analysis of the role of the phonetic label of the vowel context indicated that for most subjects, consonant categorization patterns were not influenced by the vowel label. Finally, truncated portions of natural stimuli

were identified by listeners. The roles of formant and spectral shape information indicated that for these natural stimuli, as in the case with synthetic stimuli, formant information seemed to be important for /d/'s (particularly in front vowel contexts), while spectral shape information appeared to be useful for /g/ identification. Some revised spectral shape parameters are suggested for further research.

## Acknowledgments

I would like to express my deep appreciation to the many people who have helped me through the entire 'ordeal' of this thesis. First and foremost, I thank my supervisor, Dr. T.M. Nearey, who helped me at every stage in the development of this thesis, and whose contribution to my training has been immeasurable. I was indeed fortunate to have had such a sincere friend as a supervisor. Thanks also go to his family for their kindness and hospitality. I am also deeply indebted to Dr. B. Derwing for his constant support and encouragement; his help is gratefully acknowledged. My sincere thanks to the other members of my thesis committee, Dr. J. Hogan, Dr. B. Rochet and Dr. D. Jameison for helpful suggestions and comments. I would also like to thank Dr. S. Blumstein, Dr. J. Mertus, and Dr. P. Lieberman for allowing me to use the facilities in the Phonetics Laboratory at Brown University. Thanks also to D. Cuddy, S. Rayment and S. Cohn-Sfecu of Bell-Northern Research, Ottawa, for granting me 'visiting scientist' status at the Speech and Signal Processing Laboratory. I would also like to thank the students enrolled in Linguistics 583, 1984, for gathering the raw data used in the last chapter. In addition, thanks go to all of the patient subjects who had to listen to all of my 'duck calls'. My deepest gratitude to the students and staff of the Linguistics Department for your help and support. Special thanks to 'Housemother' Tracey Derwing, and 'Colleague and Fellow Sufferer', Peter Assmann. Many thanks also go to Maureen Dow. I also express my appreciation to Mrs. Hawkes and Mrs. Cunningham who helped so much during the final stages of the thesis. Last, but not least, I thank my parents and my husband for their tremendous amount of effort, aid, and moral support. As they say in Hebrew, "KOL-HAKAVOD".

## Table of Contents

Chapter Page

1. INTRODUCTION ....................................1

1.1 The Invariance Problem ............................1

1.2 Background Review .................................5

1.2.1 Early Research at Haskins Laboratory .......5

1.2.2 Research on Cue Dependencies ..............10

1.2.3 Research on Global Acoustic Properties ....13

1.3 Models of Speech Perception ......................22

1.3.1 The Motor Theory of Speech Perception .....23

1.3.2 Feature Detection Models ..................24

1.3.3 Template or Prototype Models ..............26

1.4 Computer Recognition of Speech ...................28

1.4.1 Signal Processing and Feature Extraction ..28

1.4.2 Template or Prototype Models ..............34

1.4.3 Computer Modelling of Human Perception ....35

1.5 Overview .........................................37

2. MEASUREMENT OF FORMANT ONSETS IN VOWEL CONTEXT ......43

2.1 Description and Measurement of Data .............45

2.1.1 Speakers ..................................45

2.1.2 Data Base .................................45

2.1.3 Apparatus .................................45

2.1.4 Recording .................................46

2.1.5 Digital Gating ............................47

2.1.6 Measurement ...............................48

2.1.7 Analysis ..................................51

2.1.7.1 Relationship of Measurements to Early Perception Studies .........60

2.1.7.2 Classification ....................61

3. THE PERCEPTION OF FORMANT ONSETS IN RELATION TO F2 STEADY-STATES ....................................75

3.1 Pilot Vowel Experiment for Eastern and Western Canadian Dialects ....................................76

3.1.1 Subjects ....................................76

3.1.2 Stimuli ....................................77

3.1.3 Procedure ....................................77

3.1.4 Results and Analysis ......................78

3.2 MAIN PERCEPTION EXPERIMENT: THE ROLE OF F2v .....80

3.2.1 Subjects ....................................80

3.2.2 Stimuli ....................................80

3.2.3 Procedure ....................................83

3.2.4 Results and Analysis ......................83

3.2.4.1 Analysis of Acoustic Features .....88

3.2.5 The Use of Formant Information ............97

3.2.5.1 Summary of Formant Information ...110

3.2.6 The Use of Spectral Shape Information ....111

3.2.6.1 Stevens and Blumstein Templates ..111

3.2.6.2 Lahiri and Blumstein Ratios ......117

3.2.6.3 Kewley-Port Features .............124

3.2.7 Summary and Comparison ...................131

3.2.7.1 Conclusions .......................141

4. PHONETIC AND ACOUSTIC INTERACTIONS IN CONSONANT CATEGORIZATIONS ....................................143

4.1 22 Subjects Pooled ............................150

4.1.1 Procedure ....................................150

4.1.2 Results and Analysis ......................150

4.1.2.1 Analysis of the Effect of the Phonetic Label ....................159

4.2 Eight Subjects, Five Replications ..............163

4.2.1 Speakers ................................163

4.2.2 Stimuli and Procedure ...................163

4.2.3 Results and Analysis ....................163

4.3 Subject Differences in the Influence of Vowel Label ..........................................181

4.3.1 Subjects ................................182

4.3.2 Stimiuli ................................182

4.3.3 Procedure ...............................183

4.3.4 Results .................................184

4.4 Summary ......................................200

5. FORMANT AND SPECTRAL SHAPE CUES IN NATURAL SPEECH ..202

5.1 Onset Spectral Characteristics ................202

5.1.1 Subjects ................................202

5.1.2 Stimuli .................................202

5.1.3 Procedure ...............................203

5.1.4 Results .................................203

5.1.5 Analysis ................................205

5.1.5.1 Stevens and Blumstein Templates ..205

5.1.5.2 Kewley-Port Features .............220

5.2 Natural Data With Response Shifts ..............226

5.2.1 Subjects and Procedure ..................226

5.2.2 Stimuli .................................226

5.2.3 Results and Analysis ....................226

5.2.3.1 Analysis of Spectral Shape .......227

5.2.3.2 Analysis of Formant Information ..246

5.2.4 Summary ..................................26[illegible]

6. SUMMARY AND CONCLUSIONS ...........................265

6.1 The Role of Formant Onsets ......................265

6.2 The Role of Spectral Shape ......................267

6.3 The Role of the Vowel Label .....................268

6.4 The Effect of Vowel Context .....................268

6.5 Suggestions for Future Reseach ..................270

REFERENCES ..............................................273

Appendix I: Spectra of Natural Data; 4 msec Onset .......280

Appendix II: Spectra of Natural Data with Reponse Shifts 289

## List of Tables

Table Page

2.1 Classification Scores (Minimum Distance from /b/, /d/ and /g/ Regression Lines)....63

2.2 Overall Jacknife Classification Scores For Each Vowel....65

2.3 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /i/)....66

2.4 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /I/)....66

2.5 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /e/)....67

2.6 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /ɛ/)....67

2.7 Jacknife Classification Scores of Discriminant Function Analysis(Vowel /æ/)....68

2.8 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /ʌ/)....68

2.9 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /ɔ/)....69

2.10 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /o/)....69

2.11 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /U/)........................70

2.12 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /u/)........................70

2.13 Jacknife Classification Scores of Discriminant Function Analysis (Vowel /ɚ/)........................71

2.14 Classification Scores of Linear Discriminant Function Analysis (Vowels Pooled)..................73

2.15 Classification Scores of Second-Order Discriminant Function Analysis (Vowels Pooled)..................73

3.1 Predicted Values of F2i and F3i From Male Formant Measurements; F2v Experiment........................98

3.2 Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /o/ (F2v=900 Hz)................................99

3.3 Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel [o-U] (F2v=1050 Hz)..............................100

3.4 Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /U/ (F2v=1200 Hz)..............................101

3.5 Obtained vs. Predicted Response Categories From

Male Formant Measurements; F2v Experiment; Vowel [U-ε] (F2v=1450 Hz)....................................102

3.6 Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /ε/ (F2v=1650 Hz).....................................103

3.7 Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment Vowel /o/ (F2v=900 Hz)...112

3.8 Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment Vowel [o-U] (F2v=1050 Hz)113

3.9 Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment Vowel /U/ (F2v=1200 Hz)..114

3.10 Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment Vowel [U-ε] (F2v=1450 Hz)115

3.11 Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment Vowel /ε/ (F2v=1650 Hz)..116

3.12 Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /o/ (F2v=900 Hz)...........................118

3.13 Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel [o-U] (F2v=1050 Hz)....119

3.14 Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /U/ (F2v=1200 Hz)....120

3.15 Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel [U-ε] (F2v=1450 Hz)....121

3.16 Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /ε/ (F2v=1650 Hz)....122

3.17 Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel /o/ (F2v=900 Hz)....125

3.18 Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel [o-U] (F2v=1050 Hz)....126

3.19 Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel /U/ (F2v=1200 Hz)....127

3.20 Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel [U-ε] (F2v=1450 Hz)....128

3.21 Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz)....................................129

3.22 Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /o/.....132

3.23 Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel [o-U]...133

3.24 Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /U/.....134

3.25 Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel [U-ɛ]...135

3.26 Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /ɛ/.....136

4.1 Log-Linear Analyses; 22 Subjects Pooled; Vowel [o-U] (F2v=1050 Hz)...............................151

4.2 Log-Linear Analyses; 22 Subjects Pooled; Vowel [U-ɛ] (F2v=1450 Hz)...............................155

4.3 Log-Linear Analyses; 8 Subjects x 5 Replications (Pooled); Vowel [o-U] (F2v=1050 Hz)...............174

4.4 Log-Linear Analyses; 8 Subjects x 5 Replications (Pooled); Vowel [U-ɛ] (F2v=1450 Hz)...............177

4.5 Log-Linear Analysis of Vowel-Label Experiment......197

5.1 Correct Idntification of Natural Data..............204

5.2 Natural Data (4 msecs onset); Stevens and Blumstein Templates; Subject PFA.............................206

5.3 Natural Data (4 msecs onset); Stevens and Blumstein Templates; Subject RAH.............................207

5.4 Natural Data (4 msecs onset); Stevens and Blumstein Templates; Subject TMD.............................208

5.5 Natural Data (4 msecs onset); Stevens and Blumstein Templates; Subject MLD.............................209

5.6 Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject PFA..............................221

5.7 Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject RAH..............................222

5.8 Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject TMD..............................223

5.9 Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject MLD..............................224

5.10 Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject PFA..................228

5.11 Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject RAH..................229

5.12 Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject TMD....................230

5.13 Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject MLD....................231

5.14 Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject PFA....................234

5.15 Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject RAH....................235

5.16 Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject TMD....................236

5.17 Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject MLD....................237

5.18 Natural Data With Response Shifts; Kewley-Port Features; Subject PFA..............................238

5.19 Natural Data With Response Shifts; Kewley-Port Features; Subject RAH..............................239

5.20 Natural Data With Response Shifts; Kewley-Port Features; Subject TMD..............................240

5.21 Natural Data With Response Shifts; Kewley-Port Features; Subject MLD..............................241

5.22 Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets;

Subject PFA....................................242

5.23 Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject RAH....................................243

5.24 Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject TMD....................................244

5.25 Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject MLD....................................245

5.26 Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements; Subject PFA....247

5.27 Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements; Subject RAH....248

5.28 Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements; Subject TMD....249

5.29 Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements; Subject MLD....250

5.30 Predicted Values of F2i and F3i From Male Formant Measurements; Natural Data with Response Shifts; Speaker PFA....................................251

5.31 Predicted Values of F2i and F3i From Male Formant Measurements; Natural Data with Response Shifts;

Speaker RAH....................................252

5.32 Predicted Values of F2i and F3i From Female Formant Measurments; Natural Data with Response Shifts; Speaker TMD....................................253

5.33 Predicted Values of F2i and F3i From Female Formant Measurments; Natural Data with Response Shifts; Speaker MLD....................................254

5.34 Deviations of Predicted Values From Male Formant Measurements; Natural Data with Response Shifts; Speaker PFA....................................256

5.35 Deviations of Predicted Values From Male Formant Measurements; Natural Data with Response Shifts; Speaker RAH....................................257

5.36 Deviations of Predicted Values From Female Formant Measurements; Natural Data with Response Shifts; Speaker TMD....................................258

5.37 Deviations of Predicted Values From Female Formant Measurments; Natural Data with Response Shifts; Speaker MLD....................................259

5.38 Summary and Comparison of Obtained and Predicted Final Response Categories; Natural Data with Response Shifts; Male Speakers...................260

5.39 Summary and Comparison of Obtained and Predicted

Final Response Categories; Natural Data with Response Shifts; Female Speakers....................261

## List of Figures

2.1 Block Diagram of Digital Gating and Segmenatation...48

2.2 Male /b/ Formant Measurments........................52

2.3 Male /d/ Formant Measurements.......................53

2.4 Male /g/ Formant Measurements.......................54

2.5 Female /b/ Formant Measurements.....................55

2.6 Female /d/ Formant Measurements.....................56

2.7 Female /g/ Formant Measurements.....................57

2.8 Regression Lines for /b/, /d/ and /g/ (Male)........58

2.9 Regression Lines for /b/, /d/ and /g/ (Female)......59

3.1 Vowel Categorization of Western Canadian and Eastern Canadian Listeners.........................79

3.2 Schematic Diagram of Stimuli; F2i Experiment........83

3.3 Consonant Categorizations; F2v Experiment (22 Subjects Pooled)................................84

3.4 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /o/ (F2v=900 Hz)........................................91

3.5 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel [o-U]

(F2v=1050 Hz)....................................92

3.6 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /U/ (F2v=1200 Hz)....................................93

3.7 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel [U-ε] (F2v=1450 Hz)....................................94

3.8 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /ε/ (F2v=1650 Hz.)....................................95

4.1 F2i Marginals; Ambiguous Vowel [o-U]; 22 Subjects Pooled....................................153

4.2 F2i Marginals; Ambiguous Vowel [U-ε]; 22 Subjects Pooled....................................157

4.3 F2i Marginals; Non-ambiguous Vowels /U/ and /ε/; 22 Subjects Pooled....................................159

4.4 Syllabic Templates From Average Male Formant Measurements....................................161

4.5 Consonant Categorizations; F2v Experiment; Subject SS....................................164

4.6 Consonant Categorizations; F2v Experiment; Subject TR....................................165

4.7 Consonant Categorizations; F2v Experiment; Subject PA....166

4.8 Consonant Categorizations; F2v Experiment; Subject BC....167

4.9 Consonant Categorizations; F2v Experiment; Subject RH....168

4.10 Consonant Categorizations; F2v Experiment; Subject MM....169

4.11 Consonant Categorizations; F2v Experiment; Subject TD....170

4.12 Consonant Categorizations; F2v Experiment; Subject JS....171

4.13 F2i Marginals; Ambiguous Vowel [o-U]; 8 Subjects x 5 Repetitions....176

4.14 F2i Marginals; Ambiguous Vowel [U-ε]; 8 Subjects x 5 Repetitions....178

4.15 F2i Marginals; Non-ambiguous Vowels /U/ and /ε/; 8 Subjects x 5 Repetitions....180

4.16 Consonant Categorization For Vowel Label Experiment; Subject PA....185

4.17 Consonant Categorization For Vowel Label Experiment; Subject RH....186

4.18 Consonant Categorization For Vowel Label Experiment; Subject KT....187

4.19 Consonant Categorization For Vowel Label Experiment; Subject CL....188

4.20 Consonant Categorization For Vowel Label Experiment; Subject MD....189

4.21 Marginal of F2i; Vowel Label Experiment; Subject PA....190

4.22 Marginal of F2i; Vowel Label Experiment; Subject RH....191

4.23 Marginal of F2i; Vowel Label Experiment; Subject KT....192

4.24 Marginal of F2i; Vowel Label Experiment; Subject CL....193

4.25 Marginal of F2i; Vowel Label Experiment; Subject MD....194

# 1. INTRODUCTION

## 1.1 The Invariance Problem

Traditional linguistic theory has described the stop consonants in terms of the articulatory features of voicing, manner, and place of articulation. With the invention of the spectrograph (Koenig, Dunn and Lacey, 1946), and the Pattern Playback (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952), speech research turned in the direction of studying the acoustic output of articulation and the perception of this acoustic output.

Stop consonants are characterized by a silent (or low energy) period followed by a burst. If it is a voiceless initial stop (in English), a period of aspiration usually follows the burst; if it is a voiced stop, a voice bar is sometimes evident prior to the burst. A transition period, characterized by rapid shifts in the formant frequencies, leads into the following steady state portion of the vowel. Spectrographic studies showed that different stop consonants were characterized by differences in burst frequencies and transition movements. They also showed that transition patterns changed depending on both the initial consonant and the following vowel.

The search for invariant acoustic cues for stop consonants characterized perceptual studies in the 1950's (e.g., Cooper, et al., 1952; Hoffman, 1958). One of the first perplexing problems to arise was that locally identical phonetic events were associated with distinct phonetic percepts (or different physical events were associated with equivalent phonetic judgments). Thus, researchers could not find a one-to-one correspondence between acoustic patterns and phonetic judgments.

Researchers attempting to find invariant acoustic cues had to deal with physical variation. There are many sources for physical variation in the acoustic signal. The following are based on human performance variables:

1. Sub-threshold changes; i.e., changes in the speech signal which are undetectable by human listeners, and are produced, for example, by small changes in the speaker's articulatory gestures or environmental factors;
2. Intra-subject variation; i.e., detectable variations in the signal heard when repetitions of the same speech token are produced by the same speaker;
3. Inter-subject differences; i.e., detectable changes in the signal due to differences among speakers.

The following points (see Nearey, Hogan and Rozsypal, 1979) are examples of signal variation based on linguistic, or coding factors:

1. Phone differentiating changes; i.e., changes in the signal which cue different phonetic events;
2. Allophonic differences; i.e., language-specific acoustic differences which do not cue distinct phonological entities but are conditioned by their position within

the word (and may cue syllable or word boundaries);

3. Intrinsic allophonic changes; i.e., 'universal' (not language specific) changes in the signal due, perhaps, to inherent articulatory constraints;
4. Phone-preserving covariation; i.e., changes in the signal which are necessary in order to perceptually preserve a single phonetic event.

It is principally this last factor, phone-preserving covariation, which led to the 'problem of invariance' as defined by the Haskins group. Physical changes were necessary to preserve a perceptual constant; therefore, invariant physical events could not be found which could cue constant phonetic percepts.

In fact, the invariance problem in speech could be considered to be a type of perceptual constancy problem, since there are variable speech elements which appear to retain a constant phonetic identity (see Hochberg, 1964, for review of perceptual constancy). In addition, as with other types of perceptual constancy problems (e.g., size constancy or shape constancy in vision; Graham, 1966), *context* must be considered. For the perception of consonants, *vowel context* had to be taken into account.

There have in general been two approaches to the problem of invariance in speech perception. One approach is that separate elements (e.g, burst, formants, etc.) are initially perceived and then combined at a later stage of processing, based on 'special' unconscious inferences. This approach was taken by researchers who attributed invariance in speech perception to listeners' reconstruction of a set of underlying articulatory parameters (e.g., Studdert-Kennedy, 1976; Kuhn, 1979). They suggested that speech involves a complex coding of relatively invariant articulatory events into sound waves and therefore needs a special decoding mechanism which is different from those for general hearing. The second approach to invariance problems suggests that global properties are perceived, rather than separate elements. This approach has been taken by speech researchers who suggested that listeners attend to larger 'perceptual units', i.e., more global characteristics of speech patterns (e.g., Cole and Scott, 1974; Blumstein and Stevens, 1979; Blumstein and Stevens, 1980). They attempted to define invariant properties without reference to underlying articulatory factors.

The search for invariant physical events which cue stop consonants has motivated many investigations over the last thirty years. Early investigators found the 'sufficient' cues. Later studies investigated cue combinations and attempted to model perceptual processing.

During the 1960's and 1970's, applied problems arose which gave further impetus to the search for invariance, namely, computer synthesis and recognition of speech. If computers were to be programmed to produce speech so that human listeners could understand, the program had to supply the necessary phone-preserving cues such that humans would perceive the 'intended' message. If computers were to be programmed to understand human speech, the program had to deal with the phone-preserving covariation found in human speech. In addition, the program must be able to cope with the other variations found in natural speech. On one hand then, the computer is required to take note of the variation that is 'needed' in order to preserve a phonetic event. On the other hand, it is required to disregard certain other variations, i.e., those changes in the signal which do not affect phonetic identity, such as speaker differences or intra-speaker changes.

If it is assumed that the human listener represents a near optimal 'speech recognizer', then closer attention to the human speech perception process is a necessity. What acoustic cues are important for the human listener? How does the human listener contend with phone-preserving covariation, as well as other variation in the acoustic signal?

## 1.2 Background Review

### 1.2.1 Early Research at Haskins Laboratory

Technological advances in the 1950's provided the means to synthetically produce speech. The Pattern Playback (Cooper, et al., 1952) converted hand-painted pictures of simplified spectrograms into sound. This allowed researchers to test the perceptual effects of changing specific elements in the speech signal. In general, the aim was to isolate the acoustic patterns that acted as cues for listeners.

Cooper et al. (1952) prepared a series of synthetic stimuli by varying the frequency of the burst in front of seven different synthesized vowels. They found that high frequency bursts would be heard by listeners as /t/, while lower frequency bursts were hears as either /k/ or /p/, depending upon the following vowel. In general, a /k/ was perceived when the burst frequencies were close to the values of the second formant frequencies of the following vowel.

Their study then focused on the role of the second formant (F2) transitions. A continuum of F2 transitions was constructed whereby the initial portion gradually shifted from upward-moving to straight to downward-shifting. Listeners' identification results showed that upward rising transitions generally led to the perception of a labial consonant, while straight and downward-sloping transitions led to alveolar or velar responses. In general, straight transitions were heard as alveolar stops when followed by front vowels and velars before back vowels. Falling transitions were heard as velars before front vowels and as alveolars before back vowels. This pattern was the same for both voiced and voiceless stop consonants.

The distinction between voiced and voiceless stops could be made by synthesizing differences in the first formant (F1) transition: by cutting back the F1 transition, a voiceless stop was heard, whereas a full F1 transition was perceived as a voiced stop. Thus, F1 transitions were felt to specify manner of articulation and F2 transitions were suggested as cues to place of articulation.

The problem of invariance was clearly apparent in this first study. A burst in the frequency range of about 1500 Hz signalled a /p/ before high vowels, but cued a /k/ when followed by low vowels. Falling transitions could be perceived either as alveolars or velars, depending upon the following vowel. In other words, an identical acoustic event gave rise to different phonetic percepts depending upon the vowel context.

The authors suggested that a code of binary choices was operative in human perception of stop consonants. For example, /p/ could be perceived as the resultant of two choices - a low burst (as opposed to a high burst) followed by a rising (not falling) transition. Similarly, /t/=[+high burst], [-falling transition], and /k/=[+high burst], [+falling transition]. Even the authors admitted, however, that this could not be a realistic solution, since satisfactory stops could not be synthesized on the basis of 'high' or 'low' distinctions without regard for the exact frequency range, or degree of transition. The authors pointed to their inability to find invariance:

> [W]e - and perhaps some other workers as well - had undertaken to find the "invariants" of speech, a term which implies, at least in its simplest interpretation, a one-to-one correspondence between something half-hidden in the spectrogram and the successive phonemes of the message. It is precisely

> this kind of relationship that we do not find....(Cooper et al., 1952, p. 604)

In 1954, Schatz reported the results of a tape-splicing experiment with natural speech where the burst portions from one syllable were spliced onto the vowel portion of a second syllable. Thus, for example, the /p/ burst portion from /pi/ was exchanged with the /k/ burst portion from /ka/. She confirmed the synthetic results obtained by Cooper et al. (1952); for example, contextual vowel changes caused /k/'s from low vowels to be heard as /p/'s when appended to high vowels.

The 'locus' theory was developed in a paper by Delattre, Liberman, and Cooper (1955). The invariance was thought to be a hidden constant frequency location that was characteristic for each stop consonant place of articulation. This frequency 'locus' was a point on the frequency scale to which transitions 'pointed'; the initial portion of the transition was assumed to be missing. The authors regarded the transitions as movements from a (hidden) locus to steady state values of the following vowel. When the 'missing' part of the transition was included, however, more consonant confusions arose than without it. This was an attempt to find invariance through abstraction; a real physical invariance was still to be found.

Halle, Hughes and Radley (1957) prepared energy density spectra of the bursts and investigated the transitions via sonograms. Binary choice codes for the burst were developed, based on energy distribution patterns in the spectra. For example, the feature 'acute' was marked positive if there was significant energy in the high frequencies (above 4000 Hz); this would tend to separate /t/, /d/ and the front variants of /k/ and /g/ (i.e., followed by front vowels) from /p/, /b/ and the back variants of /k/ and /g/. The feature 'grave' was marked positive if there was significant energy in the low frequency region (500 1500 Hz). With regard to the transitions, these authors results were similar to the results of Cooper et al. (1952); they found that the regularities were complex and context-dependent.

Third formant (F3) transitions were isolated as being a cue for the perception of place of articulation in a study by Harris, Hoffman, Liberman, Delattre and Cooper (1958). In this experiment, the authors constructed continua of both F2 and F3 transitions for steady state values representing the vowels /a/ and /i/. All possible combinations of these stimuli were played for listener judgments. In general, F3 transitions shifted identification boundaries and were context-dependent. The role of F3 transitions seemed to enhance the cues provided by F2 transitions, whereas F2 transitions provided the 'major cue'. Harris et al. suggested that the F2 and F3 transition cues were independent of one another.

Hoffman (1958) studied various cues in combination. Stimuli were composed based on continua of both F2 and F3 transitions, as well as burst frequency for the steady-state value representing the vowel /a/. Synthetic stimuli contained either one cue, two cues or all three cues, in all possible combinations. He found that the burst frequency cued place of articulation in the same way for voiced stops as had been found in studies involving voiceless stops. There was also good agreement with previous studies with regard to transition cues. Hoffman suggested that the contribution of any cue was independent of other cues. In combination, he felt that adding cues was similar to addition of independent vectors.

### 1.2.2 Research on Cue Dependencies

Recent research activities have concentrated on the finer relationships of the cues that had been found in earlier studies. Dependencies and 'trading relationships' between F2, F3 and bursts were found, contrary to the conclusions of Harris et al. (1958) and Hoffman (1958), which stated that there was cue independence. Thus the search for invariance seemed even more complicated. Some of these investigations studied listener's perception when parts of the signal were missing in order to find out which of the cues were absolutely necessary.

Menon, Rao and Thorsar (1974) suggested that the cue to place of articulation was the difference between the F2 and F3 onset frequencies and the respective steady-state formants of the following vowel. They reported a measurement study which supported this hypothesis; perception tests conducted using synthetic speech also seemed to follow this pattern.

Dorman, Studdert-Kennedy and Raphael (1977) studied the relative perceptual weight of release bursts and formant transitions in an experiment with natural speech. By systematically removing these cues from syllables containing initial /b/, /d/, and /g/ with nine different vowels, they found that release bursts and transitions were 'functionally equivalent', i.e., when the perceptual weight of one increased, the other decreased. In addition, these authors investigated the functional invariance of the release burst by transposing the bursts from the CV syllables to all the other remaining vowels. Their results indicated that bursts were largely invariant, but that they were not perceptually important in a great many of the cases; only one out of 27 syllables for one speaker and 13 out of 27 syllables for another speaker were 'invariantly perceived' by subjects when the burst was transposed onto a different vowel (i.e., the burst was perceived as being the same consonant, regardless of the following vowel).

Just, Suslick, Michaels and Schockey (1978) reported a tape-splicing experiment in which consonant identifications were obtained for natural speech which had various portions of the signal spliced out. When formant transitions were removed, correct identification fell to 84%. Splicing the transitions onto other vowels caused the correct identification rate to fall to 66%. When aspiration was replaced with silence, identification fell even further. Following the Haskin's locus theory (of Delattre et al., 1955), they suggested that listeners extrapolate from the F2 transition portion to a locus.

Kuhn (1979) studied the relative influence of burst, F1, F2 and F3 transitions on the perception of place of articulation before the vowels /i/ and /a/. He produced a digitized form of the acoustic signal using parallel synthesis similar to signals produced by the analog Pattern Playback. The patterns which appeared in spectrograms, including the burst portion, were digitally modeled. He then systematically eliminated the formants and burst. His findings were:

1. Eliminating F1 (i.e., having burst + F2 + F3) lowered place indentification from 93% correct for the full digitally modeled stimuli to 90% correct;
2. Including burst and only F2 lowered identification to an average of 72%, but identification was better for consonants before /a/ than before /i/ (90% before /a/);
3. Including only F2, i.e., eliminating the burst resulted in 73% correct identification, which is virtually identical to the identification obtained when burst was present with the F2 formant transition;
4. Including burst and only F3 resulted in 57% correct identification;
5. Including only F3 (i.e., eliminating the burst) resulted in only 54% correct identification, again almost

identical to the identification obtained when the burst was present.

Kuhn also studied place identification in 'fricative' speech (speech uttered through a constricted vocal tract) which, he stated, approximates a single formant; identification fell to 86% compared to voiced speech. When bursts were removed from fricative speech, identification was at 72% correct. Kuhn found that there was a cue reweighting with regard to F2 and F3; F2 was a greater cue for point of articulation than F3 except for consonants before /i/, where F3 was greater than or equal to the F2 cue. He suggested that this pointed to the importance of 'front cavity' resonance in the perception of place (see Stevens and House, 1955, for an analysis of formant frequency and cavity affiliations).

### 1.2.3 Research on Global Acoustic Properties

During the 1970's and 1980's, some investigators were attempting to find invariant properties of the acoustic signal without reference to vowel context. They largely turned their attention to the burst portion, or to the burst and some immediately following portion of the signal.

Cole and Scott (1974) suggested that the burst and accompanying aspiration portion was the 'phantom in the phoneme', and that it represented the invariant property in the perception of stop consonants. They reported a tape-splicing experiment similar to that done by Schatz

(1954) but included the original aspiration portion of the signal. Results indicated that transposing the /b/ and /d/ bursts from /u/ to /i/ and from /i/ to /u/ did not affect identification rates. However, when the /g/ burst from /i/ was spliced onto /u/, identification plunged to 21%, and when the /g/ burst from /u/ was spliced onto /i/, identification was 82%. Winitz, Scheib and Reeds (1972) had subjects identify stop consonants from the burst portion alone, taken from syllables containing the vowels /i/, /a/ and /u/. Initial consonants were identified better than final consonants, and /t/ was identified the best (83% correct average identification for /t/, compared to 75% correct average identification for /p/ and 55% correct average identification for /k/). Identification rose when 100 msec of transition and vowel were added (77%, 83% and 70% correct average identification for /p/, /t/, and /k/, respectively.) Unfortunately, in both these studies precise segmentation criteria were not presented (see Dorman et al., 1977 for a critique).

Stevens and Blumstein (1978) created sets of synthetic stimuli with and without formant transitions. Place of articulation was consistently classified with stimuli having both a burst and transition portion, as well as with those stimuli having only the transition portion. Stimuli with only bursts and steady-states (i.e., without transitions) were not consistently classified, but the authors stressed that for these stimuli, the burst onset and the vowel onset were not 'continuous' (since no transition was present). The authors analyzed stimuli which were well identified and found that these stimuli showed invariant gross shapes of the spectrum sampled at the consonantal release (with a 25.6 msec window) for each point of articulation across all vowels. They found that these spectral patterns were evident for well-identified stimuli containing transitions only, and the spectral peaks were enhanced by the addition of the burst. These spectral patterns were as follows:

1. /b/ and /d/ had 'diffuse' spectra: there was more than one spectral peak in a limited range (1-3 KHz) and there was no prominent peak;
2. /b/ had a diffuse and falling spectrum (grave);
3. /d/ had a diffuse and rising spectrum (acute);
4. /g/ had a 'compact' spectrum where a mid-frequency prominent peak was evident; /g/'s before front vowels had this prominent peak at higher frequencies while /g/'s before back vowels had this prominent peak at lower frequencies.

Thus, Stevens and Blumstein (1978) suggested that the gross spectral shape at the 'discontinuity' (i.e., at the stimulus onset) was the 'global' invariant property in point of articulation perception.

Blumstein and Stevens (1979) measured the spectrum at the onsets and offsets of 1800 CV and VC syllables produced by six speakers. Templates for diffuse-rising, diffuse-falling, and compact stimuli were devised and compared to the spectra produced by these speakers. The diffuse-rising template correctly accepted 84% of the /d/ spectra, and correctly rejected /b/ spectra 86% of the time and /g/ spectra 88.5% of the time. The diffuse-falling template correctly accepted 82.5% of the /b/ spectra and correctly rejected 80.7% of the /b/ spectra and 90% of the /g/ spectra. The compact template correctly accepted 86.7% of the /g/'s, and correctly rejected 91.3% of /b/'s and 82.7% of /d/'s. Thus, these templates correctly accepted 85% of these spectra, although vowel contextual influences were evident.

It is worthwhile to note that the template-matching was done manually and was not computer automated. In addition, templates were fit to the spectra by trained researchers, who had experience in analyzing spectra. It would be of interest to test whether automatic application of these templates or template fitting by naive subjects would yield similar results.

In a perception experiment Blumstein and Stevens (1980) found that 10-20 msec of a synthetic CV syllable could be identified at 'above chance level' (over 30% correct identification) for place of articulation, whether F2 or higher formants contained moving or straight transitions, and whether the burst was there or not. 'Correct' identification was defined as the identification obtained when the entire synthetic signal was presented.

Identification rates *did* improve, though, with the addition of more of the moving transitions; as more moving transition was included (i.e., by increasing the duration of the signal for stimuli with moving transitions), listeners could 'correctly' identify more stimuli. Also, stimuli with moving transitions were consistently better identified than straight-transition stimuli. Thus, some transition information was found to be important.

In addition, vowel context proved to be a *very* important factor; for example, brief portions of the /bu/ synthetic stimuli had lower recognition rates than /b/'s before /a/'s or /i/'s for both moving and straight-transitions (60% recognition for the straight-transition /bu/'s vs. over 80% recognition for straight-transition /ba/'s and /bi/'s). Brief portions of /g/'s before /i/'s were recognized only slightly above 'chance level' (slightly over 35% recognition) but the results drastically improved for /g/'s before /a/ and /u/ (over 80% identification rates for both moving and straight-transition stimuli). However, the authors suggested that the 'invariant property' (i.e., gross spectral shape at stimulus onset) could be perceived with a 10-20 msec window, since 'above chance recognition' was obtained with these short stimuli. This agrees with Tekieli and Cullinan's (1979) threshold experiments, where point of articulation decisions could be made at above chance level, on the average, after only 10 msec of the stimulus. However, a claim of sufficiency cannot really be based on threshold decisions; majority responses rather than 'above chance' responses are required.

Serious difficulties arose from further consideration of the gross spectral shape of stimulus onset as the invariant cue of place of articulation. Subsequent experiments conducted by Blumstein and her colleagues tested the role of gross spectral shape in speech perception in relation to onset formant frequencies. Were listeners responding to gross spectral shape at the onset, or to onset formant frequencies as associated with a particular place of articulation and vowel environment?

Blumstein, Isaacs and Mertus (1982) constructed synthetic stimuli having appropriate formant frequencies for /b/ or /d/ in front of three vowels, /i/, /a/ and /u/. They then manipulated the shape of the onset spectrum to be diffuse rising (appropriate shape for alveolar consonants) or diffuse falling (appropriate shape for labial consonants). Their rationale was as follows:

> "If it is the case that gross shape of the onset spectrum provides invariant properties corresponding to the phonetic dimensions for place of articulation, then subjects should perceive all stimuli containing the diffuse-rising property as alveolar consonants, and all stimuli containing the diffuse-falling property as labial consonants regardless of the onset formant frequencies of the stimulus."

Results showed, however, that subjects did *not* change their phonetic labelling according to the spectral shape. With the exception of the /b/ stimulus in an /i/ context, (which was identified as /d/ when the spectrum was tilted up), manipulating spectral shape did not change listeners' categorization. Rather, onset formant frequencies determined the phonetic category to which a stimulus was assigned.

Walley and Carrell (1983) also tested the role of the global shape at stimulus onset in relation to onset formant frequencies. Using a parallel synthesizer, they altered the spectral shape of stimuli having appropriate formant onsets for /b/, /d/ and /g/ before the vowels /a/ and /u/. For stimuli with the vowel /u/, changing the onset shapes to those appropriate for the other stop categories did *not* shift perception to the other stops; rather, formant information appeared to dominate. This was also the case for stimuli with the vowel /a/, except for one case: changing the onset spectrum for /da/ to a compact shape (appropriate for /g/), while retaining the appropriate formant values for /d/, *did* change perception to /g/.

The authors also reported subject differences for stimuli with the /a/ vowel; when the onset spectra of /b/'s having appropriate formant values were altered, 10 out of 18 subjects appeared to respond on the basis of the formant cues, while only six out of the 18 subjects appeared to alter their responses according to predicted spectral onset shapes. For spectrally-altered /d/'s, 17 out of 18 subjects appeared to respond on the basis of formant cues, and for spectrally-altered /g/'s, 12 out of the 18 subjects appeared to respond to formant cues. Thus, some subects appeared to use formant-based criteria in categorization, whereas others also appeared to use onset spectral shape criteria.

A second problem for the onset spectral shape theory was found when an attempt was made to apply it to other languages. Lahiri and Blumstein (1981) studied the onset spectral properties of the phonemicaly-contrasted dental and alveolar stops in Malayalam, as well as the labial and dental stops in French. Results of the Malayalam study showed that only 57% of the dental consonants and 71% of the alveolar consonants produced by Malayalam speakers had the feature 'diffuse-rising'. In the French study, the onset spectral shape for both labials and dentals was 'diffuse-flat' (i.e., a diffuse spectral shape that was neither rising nor falling).

Lahiri and Blumstein (1981) subsequently changed their definition of the 'invariant properties' of place of articulation perception. They suggested that the *relative changes in the distribution of spectral energy to the onset of voicing* was the invariant cue. In particular, labials are characterized by either less energy in the high frequencies or a fairly even distribution of energy throughout the spectrum at stimulus onset compared to the spectral pattern at the onset of voicing. However, for dentals, there is more high frequency energy at the onset relative to the energy distribution at the onset of voicing. Thus, the new invariant factor proposed to distinguish labials and dentals was the change in the slope of the onset spectrum to the slope of the spectrum at the onset of voicing. Over 90% of Malayalam, French, and English utterances had appropriate ratio values (i.e, /d/'s had ratio values less than .5 and /b/'s had ratio values above .5).

It is important to notice that these ratios constitute a very different type of cue; what is being proposed here is *not* an 'absolute invariant' but rather two features which are taken at two points in time, and are 'combined' or integrated to form a single derived feature. Thus the contrast of later information (i.e., spectral shape at voicing onset) with earlier information (i.e., spectral shape at stimulus onset) is what is now considered important. The static nature of the features are changed, since elements of two separate points in the acoustic stream are combined as one feature.

Kewley-Port (1980) also argued against a 'static' approach in defining acoustic invariance. She noted that the Stevens and Blumstein concept of invariance was static since no temporal dimension was incorporated in their templates. Kewley-Port argued that the change in spectral distribution of energy *over time* is the appropriate cue for place of articulation in stop consonants. Using two 'time varying' features - late onset of low frequency energy and mid-frequency spectral peaks extending over time and one 'static' feature tilt of spectrum at burst onset - 88% of 3-D running spectra of CV syllables could be correctly classified. (The time-varying features separated out velars from labials and dentals.) However, this high identification rate should be compared to identification of single-syllable spectrograms to see whether the detailed representation of the 3-dimensional running spectra is, in fact, needed.

It is interesting to note that Kewley-Port (1980) suggested that a temporal dimension is required for the classification of velars, and Lahiri and Blumstein (1981) suggested that a temporal dimension is required for the classification of labials and dentals. In the attempt to find global invariant parameters, investigators were proposing features that were more complex and had to be extracted at more than one point in time.

The Steven's and Blumstein templates, the Lahiri and Blumstein ratios, and Kewley-Port's features are all *manually* fit to consonant spectra. Further tests of each of these three 'spectrally-based' parameter sets could be done on an *automatic* basis; it would be interesting to see if the same identification rates would be obtained using automatic procedures.

### 1.3 Models of Speech Perception

Several models were proposed to account for the research findings regarding stop consonant perception. The earliest models have been described above, i.e., the code and locus models that arose from the studies of the early 1950's. In the 'code' model (Halle et al., 1957), each binary choice regarding the frequency level of the burst or the transition direction was made independently of any other choices. In the 'locus' model (Delattre, et al., 1955), the listener was assumed to extrapolate from the transition onsets to a determinate locus value. Later models of stop consonant perception were really a part of general speech perception models. These were models which attempted to encompass many aspects of speech, and were thus much more general in nature.

### 1.3.1 The Motor Theory of Speech Perception

The motor theory of speech perception is a model in which speech production and perception are assumed to be inextricably linked. Since acoustic patterns showed no simple one-to-one correspondence with phonetic events, it was hypothesized that articulatory parameters had a simple correspondence with phonetic events. Perhaps, then, the acoustic patterns could first be related to articulatory parameters.

Denes (1964) outlined the basic principles of the motor theory as follows:

> The motor theory proposes that during speech recognition, we do not directly associate the sound qualities we perceive with linguistic units, the phonemes, words, etc., but that, instead, we first interpret our auditory percepts in terms of the articulatory movements needed to produce these sounds, and, in a second stage, we recognize the movements. (p.309)

In other words, this theory stated that listeners perceived stop consonants by actively reconstructing the articulatory movements needed in their production. The invariance, therefore, was the articulatory movements, or some abstraction of these movements.

The 'motor theory' position is not widely held in the current speech perception literature. Two basic questions can be asked for which there are no clear answers at present (Nearey, personal communication):

1. Can articulatory parameters be reconstructed from acoustic output without resorting to exhaustive search procedures?
2. Do there exist clearly desirable articulatory invariants that are in some significant way less context-sensitive than acoustic parameters?

### 1.3.2 Feature Detection Models

Feature detection models were proposed in the 1960's and 1970's which attempted to handle the 'invariance problem' by suggesting that variant aspects of the signal were categorized in the same manner due to specific 'feature detectors'. After the report of visual line and field detectors in the cat (Hubel and Wiesel, 1962), investigators attempted to find patterns of perception in speech that could be compatable with a theory of speech feature detectors. Eimas, Sigueland, Jusczyk and Vigorito (1971) found that repeated stimulation with a voiced stop would shift perception such that stops previously heard as being voiced would be perceived as voiceless. They attributed this phenomena to the fatiguing of 'voice' detectors and suggested that the phonetic feature of voicing as actually 'hardwired' in the brain (i.e., innate; Eimas, et al., 1971). Cooper (1974) found the same effect for place of articulation and thus suggested that there were 'place' detectors. These 'adaptation' experiments, as they are now known, started a wave of studies regarding the role and nature of 'feature detectors'. Typically, phonetic features such as place, manner and voicing were investigated.

The speech perception models which were proposed suggested that passive, largely 'hardwired' feature detectors were sensitive to some critical aspects of the signal. Two basic models were suggested, namely, those involving one level of processing, and those involving two levels of processing (see Sawusch, 1977, and Wolf, 1978, for a review). Basically, one-level models proposed that feature detectors worked solely on the acoustic signal, and therefore operated on a purely auditory level. Two-level models incorporated both an auditory and a phonetic level -- the phonetic level being an abstract 'pattern of events'.

In later experiments (Sawusch and Pisoni, 1974; Blumstein, Stevens and Nigro, 1977), a multiplicity of cues was incorporated into the feature detection system. For example, Blumstein et al. (1977) proposed that several selective detectors work on a modified input signal and that the output of these detectors, or filters, were maximized when this input signal showed certain configurations. This was a many-to-one mapping of detectors to signal, since many selective filters would give some output to a single signal input.

Feature-detection models of speech perception typically searched for particular 'simple' acoustic events (e.g., formant frequencies, VOT) that were considered in some sense 'independent' of one another in the initial extraction stage; final categorization was then modeled as some type of 'sum' of the independent contributions of these features (see Blumstein et al., 1977).

### 1.3.3 Template or Prototype Models

A more direct way to model featural non-independence was to build the dependencies into the perceptual mechanism as though there were *global* patterns of events. The difference between feature detection models and template models can be described as follows:

1. In feature-detection models, 'simple' features are extracted, independently of one another and then are combined at a 'later stage' of processing; featural non-independence is accounted for at this later stage of processing;
2. In template models, complex global features are extracted at one point in time which contain the necessary information regarding specific interactions of features

---

[1] Hochberg, 1964, describes models for visual perceptual constancies which are analoguous to feature detection (the

Investigators presented models of speech perception in which templates or prototype signals were matched with the incoming signal. Templates were considered to be 'hardwired' representations of a typical signal and if the match of the incoming signal to the template was sufficient, the signal would be classified as being a member of the type of signal that the template represented.

Massaro and Oden (1980) presented a 'fuzzy-logical' model of speech perception in which there were logical expressions regarding the attributes of a prototype or typical signal. These templates incorporated feature interactions. A decision criteron identified the incoming signal: the probability of identifying a signal to be a member of a phoneme was the degree to which the signal matched the prototype of that phoneme, compared to a match to the prototype of another phoneme.

Lieberman (1984) suggested that listeners are equipped with fully specified, syllable-sized neural templates, since experiments by Stevens and Blumstein (1978), among others, showed that perception was more highly affected by fully specified synthetic speech than by synthetic speech signals in which some signal attributes were deleted.

---

'(cont'd) structuralist approach, p. 50-60) and template models (the gestalt approach, p.74-87).

## 1.4 Computer Recognition of Speech

There are numerous studies in computer recognition of speech in which concepts from engineering, computing science, psychology and linguistics are all evident. Three basic techniques that have been employed are

1. analysis by synthesis, which was influenced by the 'motor theory' of speech perception, (Bell, Fujisaki, Heinz, Stevens and House, 1961; Paul, House and Stevens, 1964);
2. feature extraction, (Molho, 1976; Rao, 1974; Demichelis, DeMori, Laface and O'Kane, 1979); and
3. template comparisons (see below).

Most of the recent studies in computer recognition have replaced the anaysis-by-sythesis models by computationally more efficient linear predictive coding (LPC) models. LPC analysis uses correlation methods in order to model the spectral properties of the speech signal with an all-pole model. The basic idea of the method is that any given speech sample can be 'predicted' or approximated as a linear combination of past samples. There have been several formulations of LPC, including covariance, autocorrelation and maximum likelihood methods (Rabiner and Shafer, 1978).

### 1.4.1 Signal Processing and Feature Extraction

In the 1970's, computer scientists were becoming more actively involved in the problem of speech recognition (as opposed to speech synthesis problems which emphasized articulatory parameters). Thus, acoustic parameters were more fully investigated. Typically, experimenters used FFT (fast fourier transform) or LPC in order to obtain basic speech parameters such as formant frequencies, f0, and spectral and vocal tract area parameters.

Molho (1976) used a low coefficient LPC (LCLPC) in order to locate spectral peaks and compute their frequency, amplitude, amplitude rank and sharpness (related to resonance sharpness, Q). A set of binary tests was then developed which used these parameters for labelling of broad categories, such as 'burst', 'vowel', 'fricative', etc. Typical choices for the algorithm were: Is the largest peak above 3000 Hz?", or "Is the amplitude of the lowest frequency point under 3 dB?" Some choices were more complex and used rather ad hoc signal relationships, e.g., "Is the second lowest frequency peak less than twice the frequency of the lowest frequency peak minus 470 Hz?"

For stop recognition, potential bursts were identified using a low amplitude threshold. For each burst, the following parameters were extracted: spectral parameters and voicing information up to 40 msec after onset, VOT, F2 and F3 at voicing onset, amplitude change, and, as well, some contextual factors such as presence of preceeding /s/. Results indicated that stops with 'strong' bursts (/t/ and /k/, with 52% and 57% correct classification, respectively) were better recognized than those with weaker bursts (e.g., /b/, with 42% correct classification).

Specific results for /b/ /d/ and /g/ indicated that /b/ (42% correct classification, 122% false acceptance)[2] was most often misclassified as /ð/ /v/, or 'null' (i.e., consonant not detected), /d/ (33% correct classification, 66% false acceptance) was most confused with /b/ and /ð/; /g/ obtained the worst results (56% correct classification, 156% false acceptance) and was confused with various other stops and fricatives. These results indicate that, using an automatic procedure to extract these acoustic features, stop identification was rather poor.

Demichelis, DeMori, Laface and O'Kane, 1979, used relative energy within certain spectral frequency bands and ratio values of these energy levels. A 'fuzzy logic' algorithm was used which contained pre- and post-vowel context information (rules were devised which had front, central, or back vowel context). Second and third formants were also used as parameters. Unfortunately, identification results of the implementation of this algorithm were not presented. Rather, complex classification rules for /b/, /d/ and /g/ were given. It would be interesting to actually see some identification results from this classification scheme.

Rao (1974) used formant frequency and slopes of transitions; he did separate analyses on the two vowel contexts that were tested. Using principal component analysis, he suggested that slope information was 'more important' than formant frequency information. He also indicated that vowel context was not an important factor, but this claim is dubious since he did separate analyses on each vowel. In addition, only two vowel contexts (/i/ and /a/) were tested. There was no presentation of classification rates.

---

[2] More than 100% false acceptance was obtained since multiple labels were allowed.

Fujisaki and Tominaga (1982) and Kobatake and Noso (1980) also did analysis of stops for each vowel separately. Results indicate that *when the vowel context is given a priori*, classification results are very good (over 90% correct classification in both cases).

Fujisaki and Tominaga (1982) had 540 CV and VCV tokens from one speaker for the voiced stops /b/, /d/ and /g/; vowel context included /i/, /e/, /a/, /o/ and /u/. They used a measure of separability which was defined as "the ratio of inter-class variations to intra-class variations when data from two consonant classes were mapped on a common vector." Classification was then done using linear discriminant function analysis. Results show over 90% correct classification.

Kobatake and Noso (1980) used the same five vowel contexts for the voiceless stops /p/, /t/ and /k/. They had 16 male speakers record the 15 stimuli, for a total of 240 tokens. Classification was done using principal components analysis. Over 90% correct classification was obtained when the vowel was given a priori.

It should also be noted that Fujisaki and Tominaga used formant loci parameters (cf., Delattre et al., 1955), while

Kobatake and Nose used spectral information taken at four frames (frame length was 6.4 msec, frame hop was 3.2 msec). The spectra were divided into 20 bandpass filters with equal bandwidth (450 Hz), and the energy in each band was used as parameters for the principal components analysis.

Searle, Jacobsen and Rayment (1979) developed a computer model of stop consonant recognition which took into account auditory principles found in the psychoacoustic literature. They were basically concerned with the nature and role of temporal and frequency resolution in the peripheral auditory system in stop consonant perception. They allowed for good spectral resolution at low frequencies and good temporal resolution at high frequencies via 1/3 octave filters. This is an interesting aspect to test; it remains to be seen whether good temporal resolution is actually necessary at high frequencies for stop consonant resolution. Using the filter outputs, Searle et al. then employed feature detectors which detected abrupt onset, VOT, location and shape of spectral peaks near the burst and in the transition region and the average track slopes near the burst. The final decision used discriminant function analysis. An interesting problem to sort out is whether the filter system, or the features which they have selected (or a combination of both) is responsible for the good recognition results obtained by their system. It should be noted that the differences between some of these recognition schemes lie at the decision stage. Many involve 'logical decision' (or branching) algorithms (e.g., Molho, 1976), while others use statistical decision procedures such as discriminant function analysis (Searle, et al., 1979; Fujisaki and Tominaga, 1982) or classification procedures based on principle component analysis (Rao, 1974; Kobatake and Noso, 1980).

A large part of the computer recognition literature deals with 'speech understanding systems'. These systems usually have a phonetic-acoustic knowledge base coupled with higher order information bases such as phonological contraints, syntactic, semantic and even pragmatic considerations. Decisions in the acoustic stage in these systems typically use probabalistic models which yield a phonetic output that is the 'most highly probable', given the acoustic input. A major problem was that of parsing the signal such that minimal errors in phonetic labelling would occur, since an error during this labelling stage significantly decreased the overall recognition scores.

Weinstein, McCandless, Mondstein and Zue (1974), for example, in their implementation of the APEL system, initially segmented the acoustic stream into broad classes of phones (e.g., vowels, stops, fricatives, etc.) depending upon the correct detection of acoustic features such as bursts (in the case of stop detection). When a stop was detected, further spectral characterization was done on the spectral peak; the spectral peak was classified as being 'high' (between 3000-5000 Hz), 'mid' (between 2100-3000 Hz) or 'low (between 900-2100 Hz). The classification categories were /p/, /t/, front variant of /k/ (i.e., before front vowels /i/, /I/, /e/, and /ε/), and back variant of /k/ (i.e., before back and central vowels). Results indicated that 104 of 140 /t,d/'s had 'high' bursts, 9 of 15 /p,b/'s had 'mid' bursts, 32 of 36 back variant /k,g/'s had 'low' bursts, and 22 of 35 front variant /k,g/'s had 'mid' bursts. (The results are based on a 75-sentence corpus). It was evident that this classification scheme was not highly effective, since the 'mid' burst category had both velars and labials. However, the authors also suggested that 'higher-order' phonotactic constraints should be used which conceivably could increase the correct classification rates. (No results are presented, however, for improvement in identification with addition of such factors).

Other investigators looked at different ways of computer speech recognition which used the notion of templates or stored prototypes in an attempt to increase overall correct recognition rates.

### 1.4.2 Template or Prototype Models

Recognition systems which use templates have a stored 'prototypical' speech element which is compared to the incoming signal; if a high enough match is made between the signal and the template, the signal is labelled as being a member of the class which is represented by the template. The matching procedure usually incorporates dynamic programming and other normalizing methods so that minimal error is obtained.

Templates have been devised for transemes or diphones (the dynamic portion between two successive segments) by Paul and Rabinowitz, 1974; Silverman and Dixon, 1976; Mariani and Lienard, 1977; Klatt, 1980; and Schwartz, Klovstad, Makhoul and Sorenson, 1980. They have also been developed for syllables (Fujimura, 1974) and words (Klatt, 1980; and Bridle and Sedgwick, 1977). It is interesting to note that the transeme concept was also mentioned in the psychology literature (Wickelgren, 1969).

A current controversy in using template models involves the *size* of the templates, and the features which are incorporated into each template. Should templates be based on smaller 'units' such as the diphone, or should larger sized templates, such as syllabic-size or word-size templates be used? Most current practical systems use word-size templates, but this has the obvious drawback of storage limitations and long search routines.

### 1.4.3 Computer Modelling of Human Perception

One important but rather neglected area in computer research is that of modelling human speech perception. Most of the computer models to date have dealt with lower auditory functions and basic hearing models. These models have tried to examine and explain various psychoacoustic phenomena by looking at how different computer hearing models behave. Auditory models have been developed by Chistovich, Kozhevnikov, Lesogor, Shuplijakov, Taljasin and Tjulkov (1974), Dolmazon, Bastet and Shupljakov (1977), and Zwicker, Terhardt and Paulus (1979). Searle et al. (1979) also used aspects of hearing in classification of stop consonants.

Recently, some investigators have tried to model human speech perception, although some models show heavy influence from the computer recognition field. Klatt's (1980) model, for example, was composed of two parts: SCRIBER, which uses diphone templates, and LAFS, which generates lexical hypotheses directly from the acoustic input (i.e., a word recognizer).

Edwards (1981) modeled the perception of stop consonants as a set of probabalistic vectors based on categorization functions obtained from measured signal parameters. Ten voicing parameters and seven place of articulation parameters were assessed for their relative contribution in classifying the stop categories. In order to reflect the interaction of place and voicing factors in perception studies, Edwards separated the voicing and place of articulation decisions, but incorporated place of articulation features in the voicing decision, and voicing features in the place of articulation decision. The magnitude of his vectors reflected the probability of successful identifications. These were determined by looking at the relative separability provided by each parameter. For example, a particular VOT which was sometimes evident for voiced stops and sometimes evident for voiceless stops would not be given the same weight (i.e., the same vector magnitude) as a VOT that was always categorized as having the same voicing. Edward's models obtained recognition scores similar to those of human listeners. His voicing model, in fact performed as well as trained human listeners. His main idea was that listeners must be responding to *several* parameters when identifying stop consonants and, as well, these parameters interacted in specific ways. Acoustic redundancy, he maintains, is necessary for speech perception and should therefore be modeled in computer speech recognition programs.

The use of human listener categorization results (as opposed to separability of categories based on measured acoustic properties) would be an important addition to the computer speech recognition field and an approach which merits further investigation. (See Nearey and Hogan, in press, for additional models of categorization based on the relationship of categorization and speech measurements.)

## 1.5 Overview

To understand speech perception, several major subjects must be invesigated:

1. What *features* are relevant and how are these features processed by the human perception system? Do listeners attend to gross spectral patterns, or to relations

between spectral patterns at various points along the acoustic stream? Perhaps formant onset information or a relationship of onset information to the following vowel is important. Another possibility is that listeners attend to both spectral shape *and* formant information.

2. What is the role of *context* (in particular, vowel context) in listeners' perception (and speakers' production) of stop consonants? To date, most research has dealt with stop consonants in only very limited contexts, namely, before vowels at the extreme ranges of the vowel triangle (i.e., /i/, /a/, and /u/). What happens in the case of other vowels, or in the case of 'ambiguous' vowels? To what degree, if any, do consonant and vowel identification interact?
3. If context is an important factor, what is the *level of processing* by which listeners are affected by this context? Do listeners operate at an *auditory* level, whereby only acoustic information of the context is used, or do they operate at a level whereby *phonetic labelling* of the context is also important?
4. What is the *size of the perceptual unit*? Do subjects respond to segment-sized units or syllable-sized units?

In this research study, the *processing level* of the perceptual unit is addressed. Two other interesting questions, however, are the following:

1. What is the *nature* of these perceptual units? Do they have features which interact, or are their features

'independent' in the sense described by Hoffman (1958)?

2. How are these perceptual units actually *used* during the speech perception process? Do features need to be extracted at two or more points in time and then later 'combined' in some way, or do listeners attend to 'global invariants'? Perhaps reaction time experiments would shed some light on the timing factors of feature extraction. Although this is an interesting problem, it is not addressed in this research project.

In this thesis, several investigations are presented which test the production and perception of voiced stop consonants in vowel context. The following questions are addressed:

1. How does vowel context affect the production of stop consonants? Do male and female formant measurement data show similar patterns?
2. How does vowel context affect the perception of stop consonants. What part of the vowel is involved (e.g. F2 or F3)?
3. What is the level of processing involved if context is important? Are subjects influenced only by an auditory level of the context or do the contextual phonetic labels play a role? Are there subject differences involved?
4. What acoustic features are important for stop consonant place of articulation perception? Are these features vowel context-dependent, or vowel context-independent?

5. What is the perceptual 'window length' for stop consonant perception? How much of the signal is required for correct identification? Is this dependent on the following vowel, or is it context-independent?

In the first study, measurements were taken of CV syllables covering a wide range of English vowels, using both male and female speakers. Formant onsets in relation to the vowel steady-states were measured.

In the perception studies using cascade synthetic speech, formant onsets were systemmatically manipulated in front of several different steady-state portions. In spite of recent interest in the role of global spectral shape, there is ample experimental evidence from early research that *local* spectral properties (i.e, formant peak frequencies) have a strong influence on consonant perception. Many recent studies have manipulated the global shape of the spectrum at stimulus onset, which includes the initial burst. However, perception studies revealed that the onset shape characteristics could not be the sole cue for consonant identification. It seems desirable, therefore, to extend some of the previous research on formant continua with modern synthesis techniques. Since onset frequencies interact in a complex way to produce spectral shape in cascade synthesis (cf., Klatt, 1980) it is of interest to test the roles of formant information and onset spectral information for these synthetic stimuli.

The particular vowel contexts that were chosen covered the range represented by /o/-/U/-/ɛ/. They were chosen for the following reasons:

1. These vowels cover the range of the invariance problem for F2 discussed in this chapter. Early research showed that straight transitions were heard as alveolars when followed by *front* vowels (i.e., /ɛ/ in our case) and as velars when followed by *back* vowels (i.e., /o/ and /U/) while falling transitions were heard as velars before front vowels and as alveolars before back vowels.
2. These vowels are not at the extremes of the vowel triangle for which most of the current research has focused. It is of interest to test whether results obtained for the vowels at the extremes of the vowel triangle replicate with these vowels.
3. The vowels /o/, /U/, and /ɛ/ lie along an F2 continuum and hence only one factor of the vowel needed to be manipulated, in contrast to other studies in which frequency and temporal factors were manipulated differently, depending on the vowel context.

The continuum of steady states included two 'ambiguous' vowels ([o-U] and [U-ɛ]) in order to test the influence of vowel labelling on consonant identification. How is the same acoustic information treated by listeners who identify the vowel portions differently?

In a perception study using natural speech, brief portions at the beginning of of the CV signal, varying in duration from 12 msec to 40 msec were presented to listeners. An analysis was done for those cases where perception shifted from one stop category to another. The role of vowel context is also discussed.

The general findings of this research project was that vowel context plays a considerable role in stop consonant perception. In addition, it was found that both formant and spectral shape information could be interpreted as being important for perception of both synthetic and natural speech tokens. The phonetic label of the vowel context was found to have only a minor and inconsistent effect on stop consonant categorization patterns.

# 2. MEASUREMENT OF FORMANT ONSETS IN VOWEL CONTEXT

This chapter describes a measurement study of formant onsets in vowel context. Previous measurements have been done using the sound spectrograph (Ohman, 1966; Menon, Rao and Thorsar, 1974; Fant, 1973). Only one other study, to date, has measured formant information with digital signal processing techniques (Kewley-Port, 1980), and this particular study used only one male speaker. It was the purpose of the present experiment to measure formant onsets and the formants of the steady-state portions of the following vowels of speech utterances spoken by five males and five females using digital signal processing techniques. Thus, in contrast to other measurement studies, this study investigated subject variability using both male and female speakers. A larger range of vowels was also investigated, including the vowel /ɚ/, for which F3 steady-state changes are evident. This large range of vowels maximally covered the transition invariance problem described in Chapter 1.

Preliminary analysis by Nearey (personal communication; see also Nearey and Shammass, in preparation) of data from the literature indicated that F2 and F3 formant onsets and steady states were approximately linearly related for each consonant, /b/, /d/ and /g/. Nearey also suggested that such relationships could be used in the classification of stop consonants and perhaps have some relation to listeners' categorizations of synthetic stops in earlier experiments. The present experiment is an attempt to test and refine these hypotheses.

In contrast to previous work, which attempted to correctly classify stop consonants based on either formant trajectory directions or loci of formant trajectories, this study attempted to incorporate vowel information into the analysis and thus tried to classify stops with combined information on formant onsets and formants of the steady-state portions of the following vowel. This approach was also motivated by references in the literature that if the vowel context was known, stop consonants could be accurately classified according to place of articulation (Kewley-Port, 1980; Kobatake and Noso, 1980).

In this study, consonant categorization was based on onsets and steady-states of the second and third formants of the CV syllable. By using this limited amount of information, is it possible to get high identification rates? Is it possible to correctly classify consonants based on only some aspects of the following vowel (i.e., F2 and F3 steady-states), while disregarding other vowel information such as F1, formants higher than F3, offglide, vowel duration, and vowel label? How well would the classification scheme work on both males and females? Finally, is there an overall pattern of the relationship between formant onsets and the steady-state portions of the vowel? Is the invariance of the formant trajectories explainable in terms of a pattern based on vowel formants? This study attempted to provide an overall framework for unifying measurements of both vowel steady-states and formant onsets.

## 2.1 Description and Measurement of Data

### 2.1.1 Speakers

Ten speakers, five male and five female, were recorded. All subjects were native speakers of Canadian English and had no history of speech or hearing disabilities. All subjects were phonetically trained graduate students in linguistics.

### 2.1.2 Data Base

A written list of 42 CVC syllables was provided for each speaker, where the first consonant was either /b/, /d/ or /g/, the vowel was one of /i, I, e, ɛ, æ, ʌ, ɔ, o, U, u, ɚ, aI, aU, oI/, and the last consonant was /d/. The syllables with diphthongs were placed at the bottom of each of the /b/, /d/, and /g/ lists to control for list intonation effects and were not analyzed. The list of syllables was written in IPA, and examples of comparable English CVC words were also provided.

### 2.1.3 Apparatus

The instruments below were used in this study. Their technical specifications follow.

1. Microphone: Sennheiser MD 42IN, frequency response 30-17000 Hz 5 dB; sensitivity .2 mV/microbar at 1000 Hz;

cardioid directionality, using the 'S' (speech) setting which slightly raises the relative amplitude of the higher frequencies.

2. Tape Recorder: TEAC A-7030, frequency response 50-1500 Hz 2 dB; speed 7.5 ips., SNR 58 dB.
3. Audio-frequency Filter: Frokjauer-Jensen type 400, frequency response slope 36 dB/oct.
4. Minicomputer: PDP-12A; word length 12 bits; A/D, D/A converters 10 bits; operating systems OS/8 and Alligator.[3]

### 2.1.4 Recording

Subjects were individually recorded in a sound-insulated recording room. In order to eliminate possible crosstalk effects, only the left channel of the TEAC was used. In order to regulate the tempo of speaking, each syllable was first presented to the subject from a master tape on which a 'master' speaker was recorded. The tempo of the master tape was made constant by having the master speaker utter the syllables at the same rate as digitally prepared beats. The master tape was presented with a Sony tape recorder over Sony headphones.

Subjects were asked to repeat the syllables at the same rate as the presentation on the master tape. Amplitude levels were carefully monitored to avoid distortion. Three repetitions of the syllable list were recorded. Only the last two repetitions were analyzed.

---

[3] The Alligator programming system, developed by Stevenson and Stephens (1978) is written in OS/8 PAL 12D assembly language and is designed for psychoacoustic experimentation. The system is executable on PDP-12 computers.

### 2.1.5 Digital Gating

Digitization was done by an interactive Alligator program. For each CVC syllable, the initial burst portion of the first consonant to the silent gap of the final /d/ was digitized and stored on tape. Special care was taken to position the initial cursor as close as possible to the actual burst, so as not to store any prior silent portion. The audio signal from the tape recorder was bandpass filtered (68-6800 Hz) in order to eliminate 60 Hz hum and possible speech components above 8 kHz before digitizing the signal. Care was taken to avoid signal clipping, while still maintaining the broadest possible range of quantization. The wiring diagram is shown in Figure 2.1.

### 2.1.6 Measurement

Formant candidates were calcaluated using programs written by T. M. Nearey and were based on linear prediction algorithms (Markel and Grey, 1976). The first 90 msec of each syllable were analyzed every five msecs. For some subjects, particularly for female speakers, the analysis was redone with 2.5 msec hops. The sampling rate was at 16,000 b.p.s. and the analysis window was set at 256 points (16 msec window).

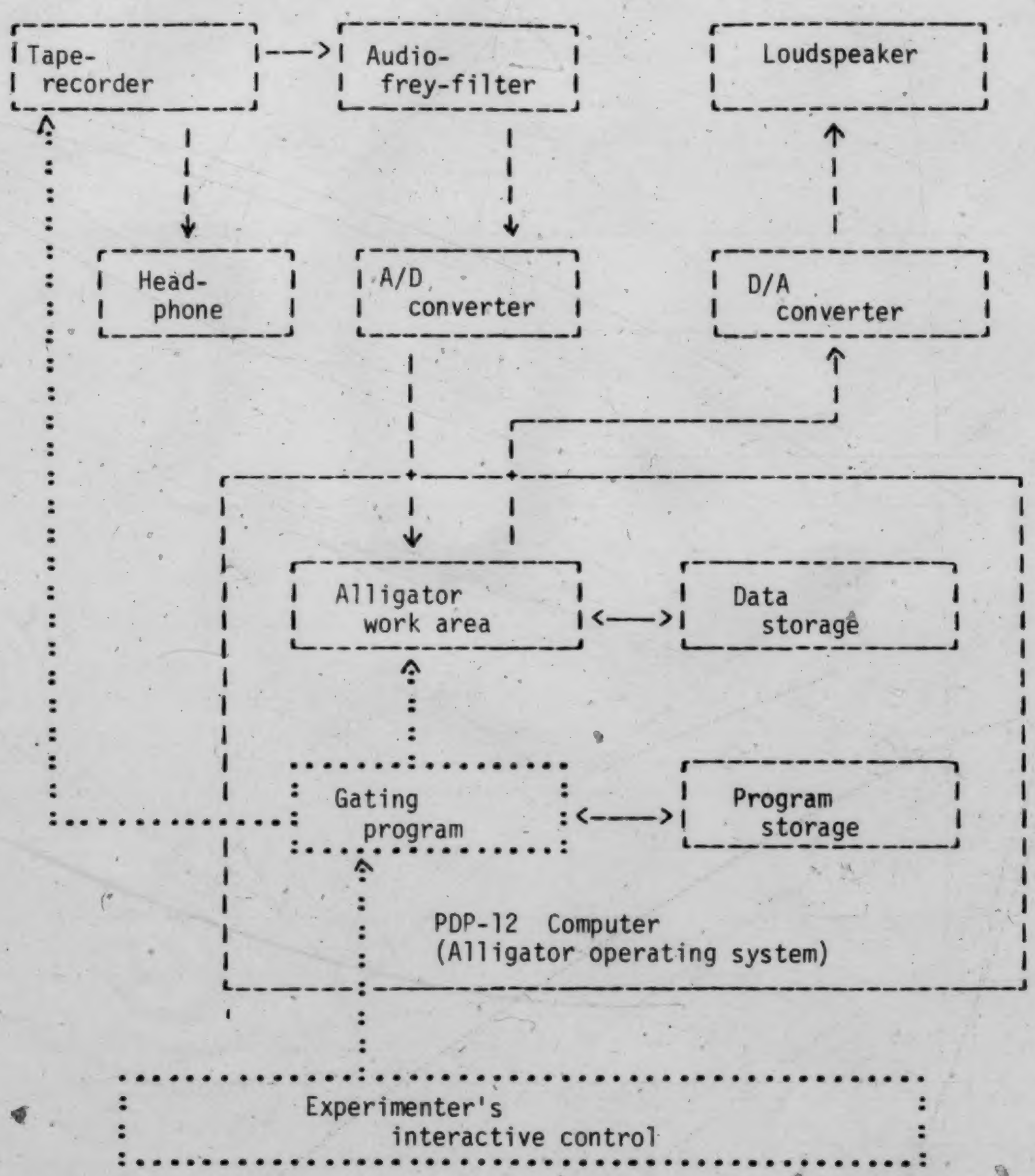

Tape-recorder
Audio-frey-filter
Loudspeaker
Head-phone
A/D converter
D/A converter
Alligator work area
Data storage
Gating program
Program storage
PDP-12 Computer
(Alligator operating system)
Experimenter's interactive control

FIGURE 2.1 Block Diagram of Digital Gating and Segmentation

* Solid Arrows indicate Signal Flows; Dotted Arrows, Control Flows; Solid Boxes, Devices; and, Dotted Boxes, Controllers.

A triangular lag window in the frequency domain was incorporated into the program in order to smooth the spectral shape (Tohkura, Itakura and Hashimoto, 1978). This was particularly important in the analysis of female speech, since spurious peaks, resulting from harmonic components of the high fundamental, could otherwise be incorrectly chosen as formant candidates. LPC analysis with 20 predictor coefficients was used to estimate the spectra of male speech, while the spectral estimate of female speech required 16-18 LPC co-efficients.

Formant candidates were determined using a method described by Christensen, Strong and Palmer (1976). The program provided frequencies and amplitudes of the formant candidates. By judiciously choosing the number of coefficients and the spectral smoothing factor, formant trajectories could be readily tracked by hand. The F2 and F3 candidates were manually tracked based on the following criteria:

1. The preceeding and following formant candidates were within 50 Hz of the present formant candidate;
2. A definite trend of formant values was evident (i.e. either rising, falling, or constant);
3. Amplitude values were at fairly high levels (20 dB or above for F2, 15 dB or above for F3).

Occasionally, the program would miss a formant at a particular slot. For example, sometimes the F1 track would miss F1 and put F2 in the F1 slot. In this case, the F2 slot was filled by the F3 candidate, and the F3 slot was filled by the F4 candidate. This occurred for F1, F2 and F3 tracks, but in almost all cases the appropriate formant candidate was recoverable by looking at the neighboring formant tracks (see Lennig, 1978).

In the case of female speech, spurious candidates (i.e., peaks not corresponding to formants) were found, even after much effort in choosing the appropriate number of LPC coefficients and the appropriate spectral smoothing factors. In such cases, the analysis was redone with 2.5 msec hops, where spurious peaks could be spotted more easily, since they are generally less stable than true formants over time. The formant candidate was picked according to the aforementioned criteria, disregarding the spurious formant candidate.

Four measures were selected manually and recorded for each stimulus: F2 onset, F3 onset, F2 steady-state, and F3 steady-state. The formant onset values were defined as the first well-tracked formant candidate having an RMS value of 20 or more for F2, and 15 or more for F3. The steady-state values were defined as the values 60 msec after formant onset, provided that neighboring formant candidates were within a 50 Hz range and a fairly constant formant track was evident.

### 2.1.7 Analysis

Male and female data were analyzed separately. The values of F2 onsets (F2i) were plotted against F2 steady-states (F2v), and F3 onsets (F3i) were plotted vs. F3 steady-states (F3v), for each consonant /b/, /d/ and /g/. Figures 2.2 - 2.7 show the resultant graphs. Figures 2.2, 2.3 and 2.4 represent male data for /b/, /d/, /g/, respectively. Figures 2.5, 2.6 and 2.7 show female data for consonants /b/, /d/, /g/, respectively. (Letter symbols refer to vowel category, with A=/i/, B=/I/, C=/e/, D=/ɛ/, E=/æ/, F=/ʌ/ G=/ɔ/, H=/o/, I=/U/, J=/u/ and K=/ɚ/).

A striking linear relationship is seen in these graphs. The /b/ plots (Figures 2.2 and 2.5) show very little scatter around the regression lines, while /d/ graphs (Figures 2.3 and 2.6) show rather large variance, as do the /g/ graphs (Figures 2.4 and 2.7). However, the F3i vs. F3v plots for /b/'s (in both the male and female data) shows a conspicuous gap of /b/'s for F3v less than 2400 Hz and F3i less than 2200 Hz; in fact, the only /b/'s having lower F3i and F3v values are /b/'s before the vowel /ɚ/, which has a lower F3 steady-state.

The data was analyzed by a robust linear regression procedure (Velleman and Hoaglin, 1981) in order to reduce the effect of outliers. The linear regression lines which were fit to the data, are shown in Figure 2.8 (male) and Figure 2.9 (female). Both the male and female data follow very similar trends. The F2i vs. F2v plots show that at low

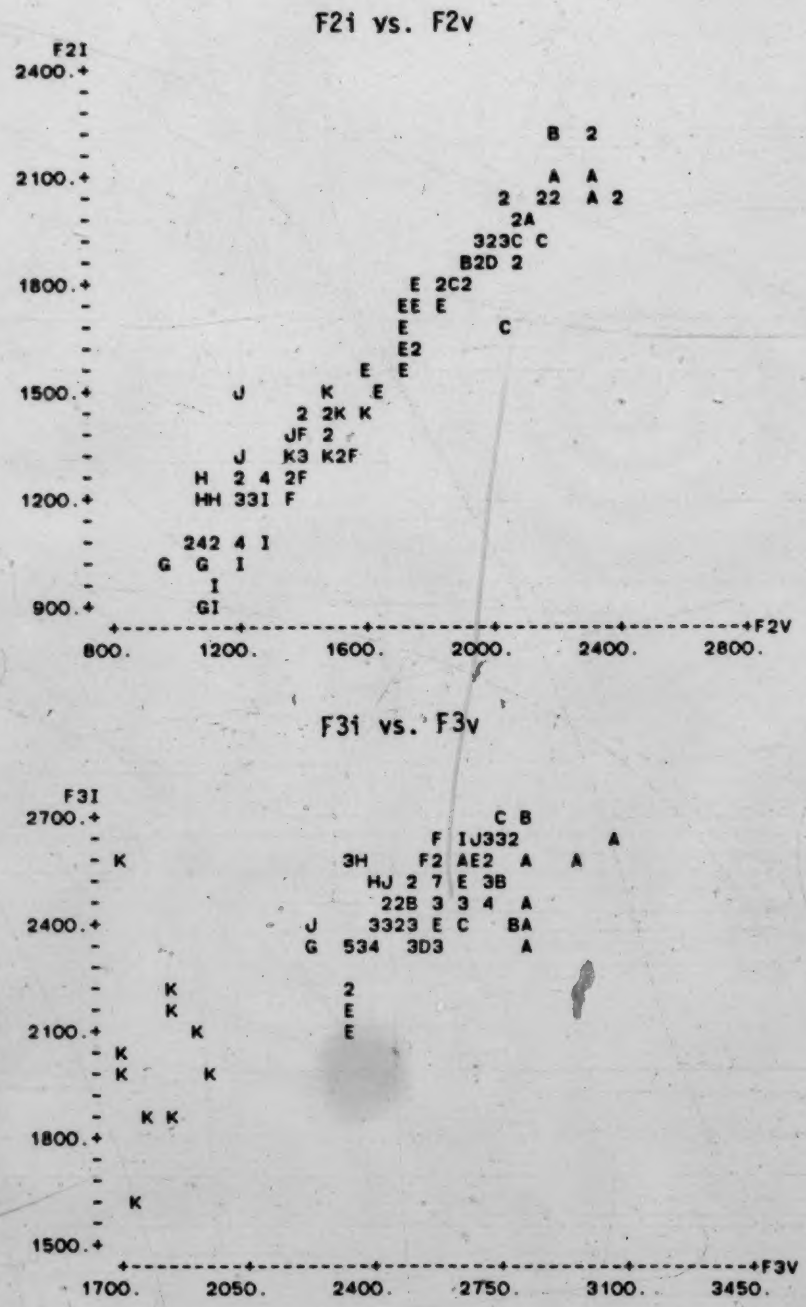


FIGURE 2.2 Male /b/ Formant Measurements

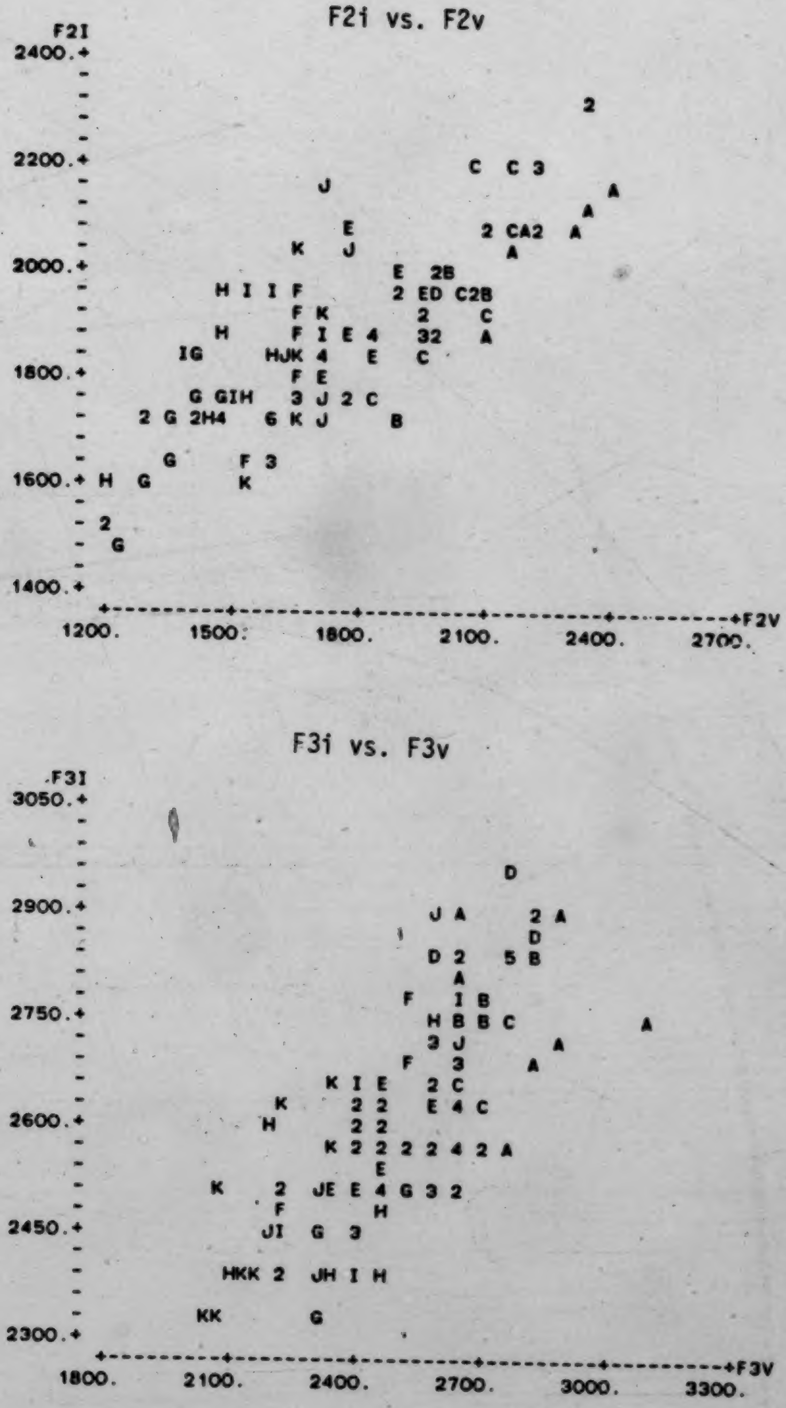
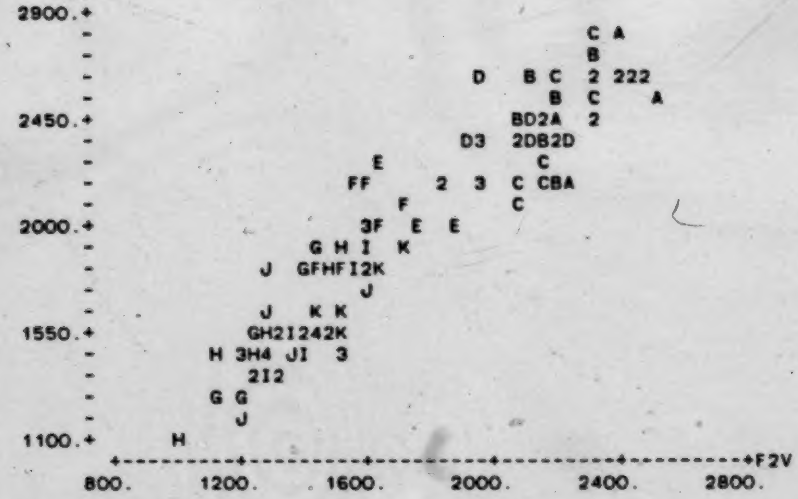


FIGURE 2.3 Male /d/ Formant Measurements


F2i vs. F2v

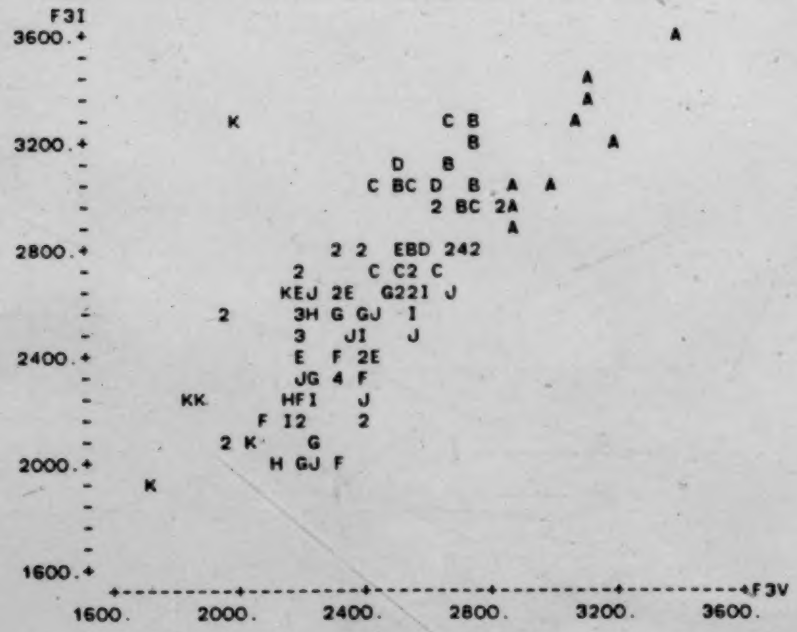
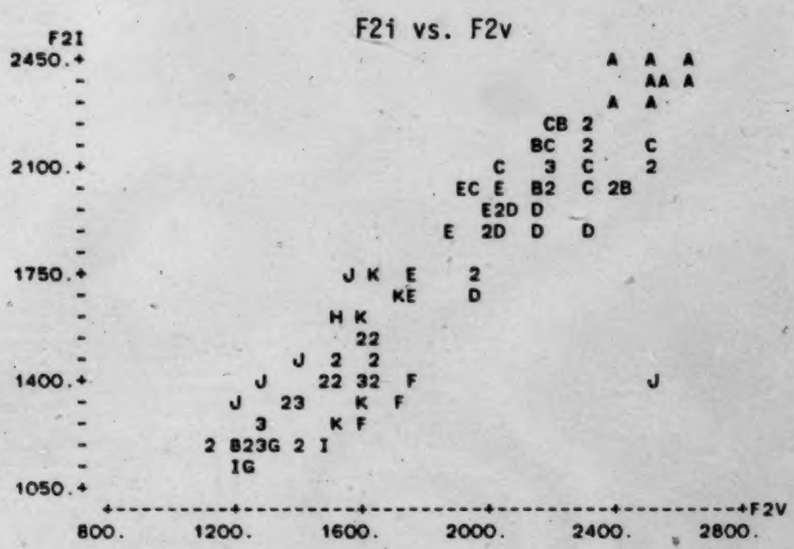
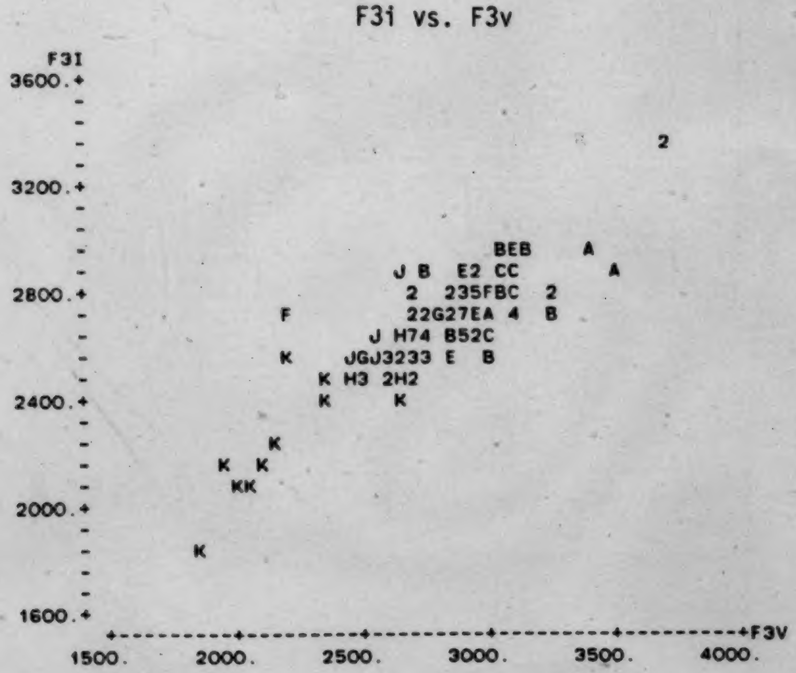

F3i vs. F3v

FIGURE 2.4 Male /g/ Formant Measurements

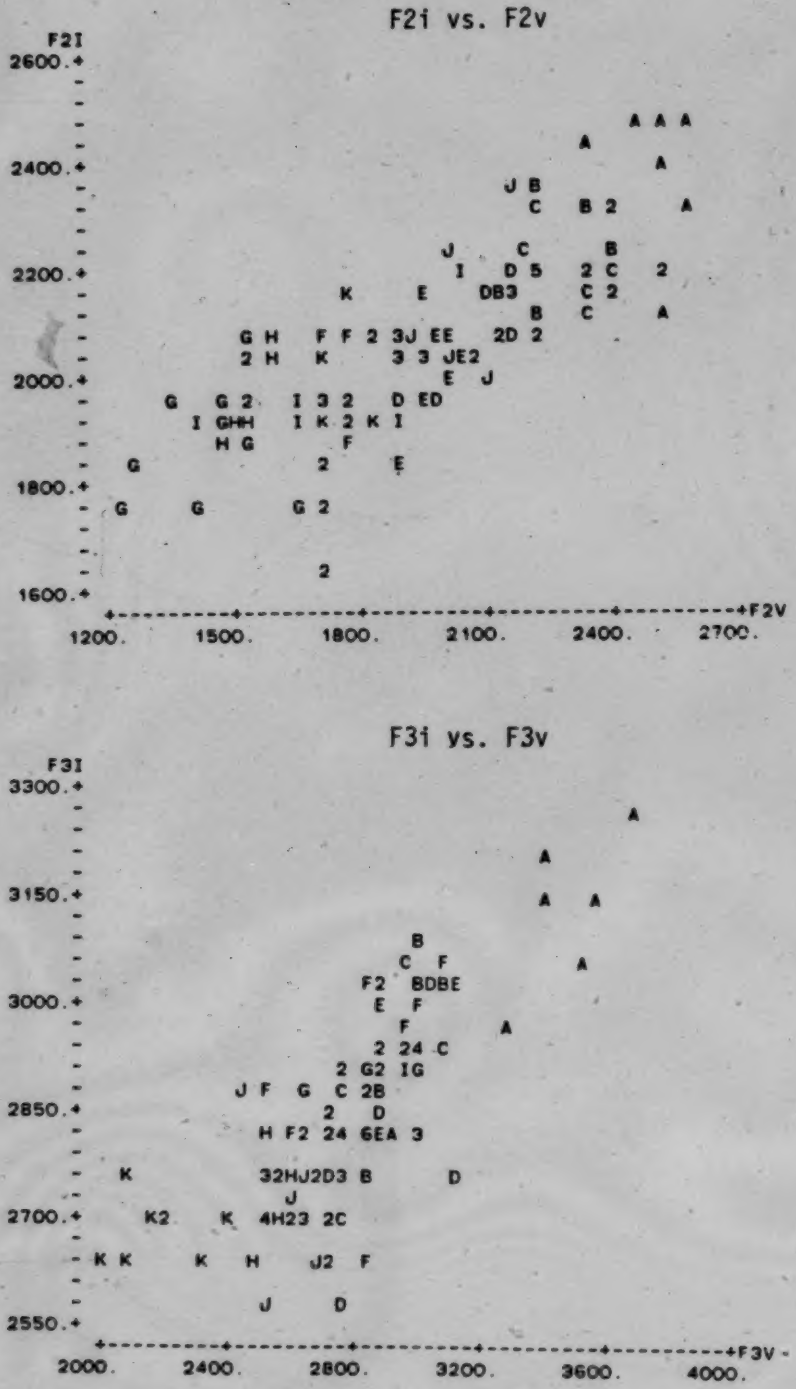


F2i vs. F2v



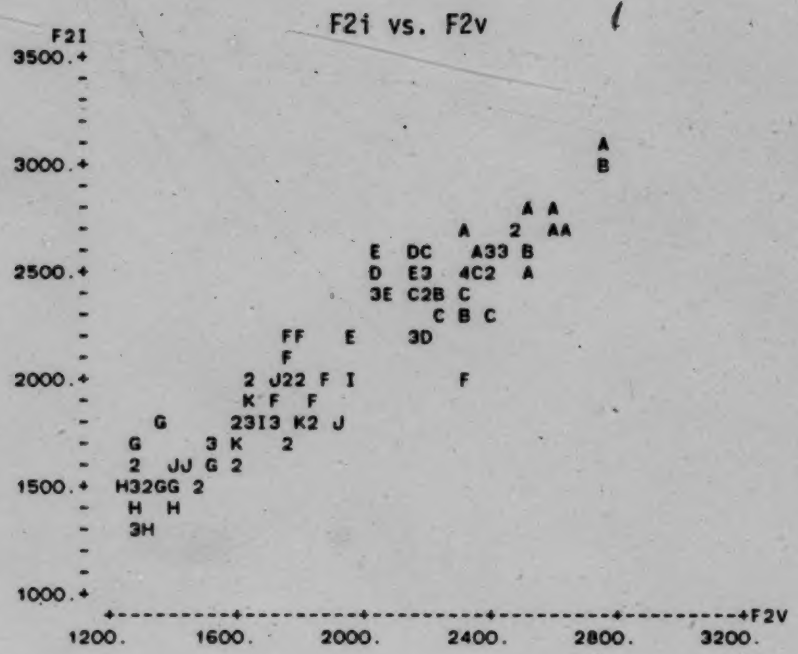
F3i vs. F3v

FIGURE 2.5 Female /b/ Formant Measurements


F2i vs. F2v

F3i vs. F3v

FIGURE 2.6 Female /d/ Formant Measurements


F2i vs. F2v

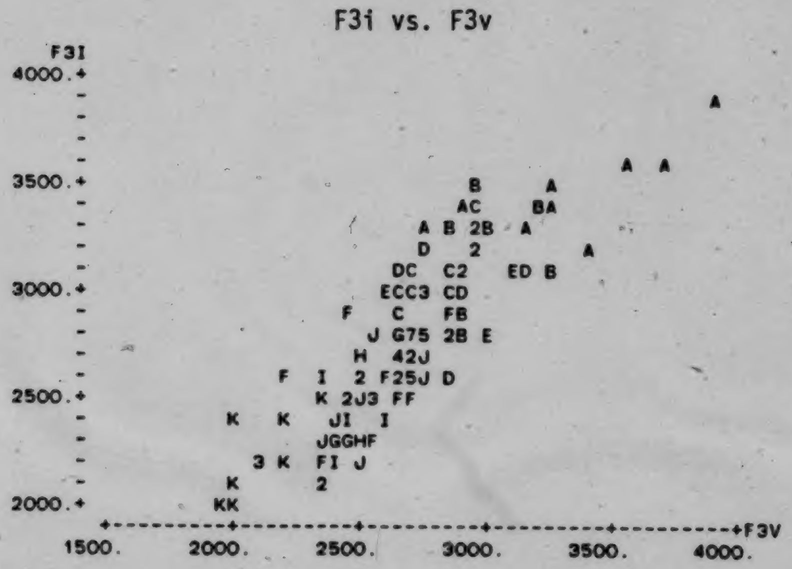

F3i vs. F3v

FIGURE 2.7 Female /g/ Formant Measurements

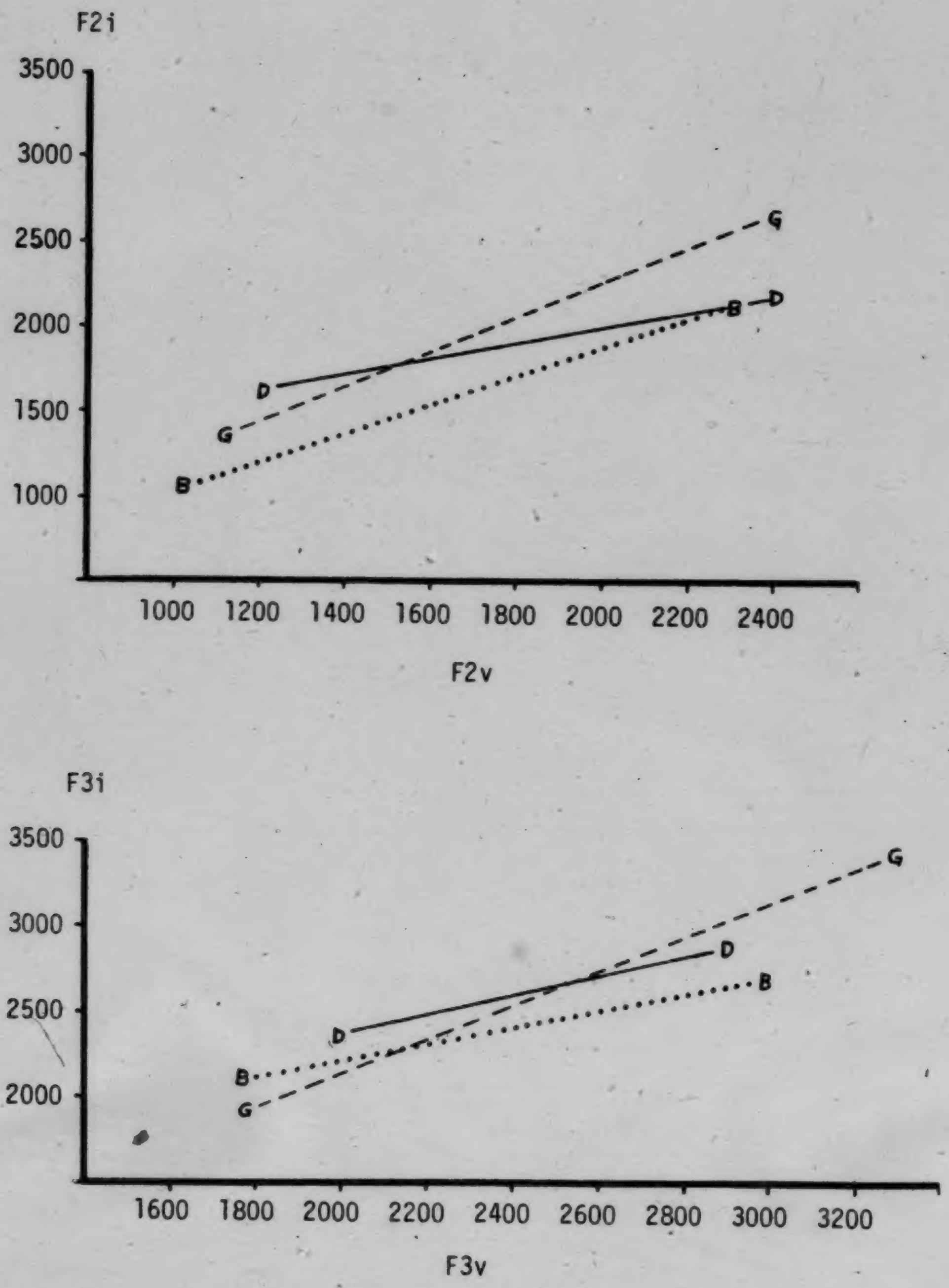


FIGURE 2.8 Regression Lines for /b/, /d/ and /g/ (Male)

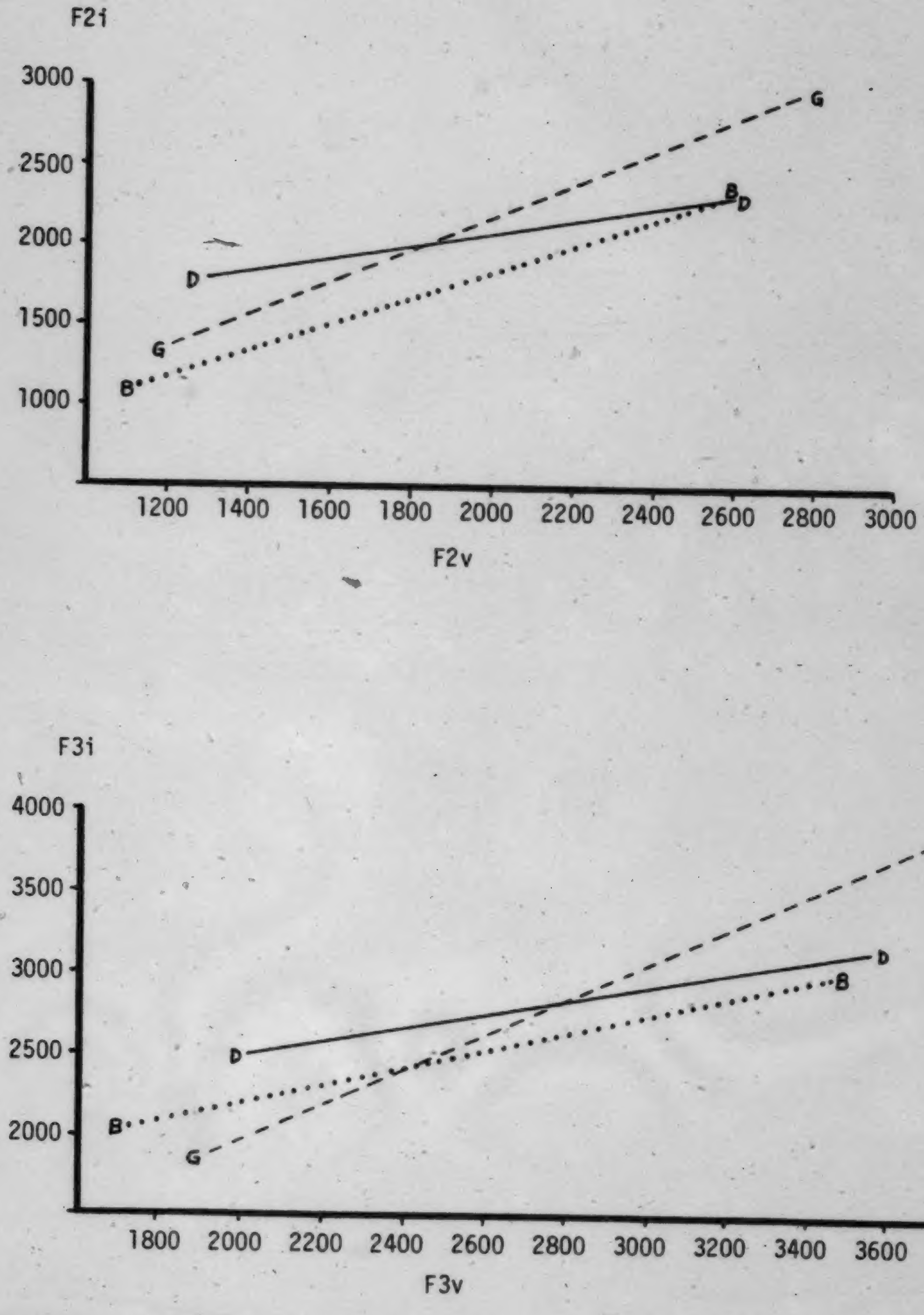


FIGURE 2.9 Regression Lines for /b/, /d/ and /g/ (Female)

F2 steady-states, the F2i of /b/ are lower than F2i of /g/ or /d/, and F2i of /g/ are lower than F2i of /d/. However, at higher F2v values, the formant onsets of /b/ and /d/ are nearly equal, while both are well below the F2i of /g/. The 'cutoff point' when F2i of /g/ is higher than F2i of /d/ is at an F2v value of approximately 1550 Hz for males and 1850 Hz for females, and is represented by the intersection of the /d/ and /g/ lines.

There is a striking similarity of patterns between the male and female data. In fact, male and female patterns fit nicely together if adjustments for normalization (to approximately 20% higher values for female formant measurements) are considered (see Nearey, 1978).

#### 2.1.7.1 Relationship of Measurements to Early Perception Studies

The measurements seem to be in general accord with the early perception experiments of formant transitions (Cooper, et al. 1952). These studies showed that rising transitions were identified as /b/'s; here, /b/'s have formant onsets lower than their corresponding formant steady-states. They also showed a change in the '/d/-/g/ crossover', which corresponds to the /d/-/g/ crossover patterns in the present measurements. Before front vowels (i.e., high F2v's), falling transitions were perceived as velars while straight transitions were perceived as alveolars; here, at high F2v's the F2i of /d/ are lower than the F2i of /g/ (i.e., /g/'s have falling transitions while /d/'s have straighter transitions). Before back vowels (i.e., low F2v's), falling transitions were perceived as alveolars, while straight transitions were perceived as velars; here, at low F2v's, F2i of /g/ are lower than than F2i of /d/ (i.e., /d/'s have falling transitions while /g/'s have straighter transitions).

The F3i vs. F3v graphs also show similar trends for both males and females. For low F3 steady-states, /g/ onsets are lower than /b/ or /d/ onsets, while /b/ onsets are lower than /d/ onsets. At mid-frequency F3 steady-states, /b/ onsets are lower than /g/ or /d/ onsets, and /g/ onsets are lower than /d/ onsets. At high frequency F3 steady-states, /b/ onsets are lowest, while /d/ onsets are lower than /g/ onsets. The 'cutoff' steady-state frequencies at which /g/ onsets are lower than /b/ onsets are approximately 2,150 Hz for males, and 2,450 Hz for females, as represented by the intersection of the /b/ and /g/ lines. The 'cutoff' steady-state values at which F3i of /d/ are lower than F3i of /g/ are approximately 2,550 Hz for males and 2,775 Hz for females, as represented by the intersection of the /d/ and /g/ lines.

### 2.1.7.2 Classification

In this section, the classification of stop consonants based on the regression lines of Figure 2.8 is discussed. Previous investigators have classified stops in terms of F2 and F3 transitions (e.g., Rao, 1974). Note that neither F2 nor F3 trajectories conditioned alose are sufficient to specify the stops uniquely, since the regression lines for the three consonants intersect, indicating category overlap (cf. Kewley-Port, 1980).

The regression lines of the F2 and F3 trajectories were used in a classification algorithm based on the squared sum of the overall minimum distances from the F2i vs. F2v and F3i vs. F3v regression lines. Using the average male and female coefficients, the entire data set (of 220 tokens) was correctly classified 73.9% of the time. Two cross-validation classifications were done. Using the male coefficients, the female data set was correctly classified 69.1% of the time; using the female coefficients, the male data set was correctly classified 64.8% of the time. The breakdown of correct classification according to consonant class (/b/, /d/ or /g/) is given in Table 2.1.

This measurement study showed a definite linear pattern of formant onsets vs. formant steady-states. Furthermore, this pattern was evident for both male and female speakers. The classification results show that 73.9% correct classification, for both male and female speech, can be obtained using these parameters. Formant onsets can be viewed as a linear pattern in relation to the following vowel. It is important to note that no

## TABLE 2.1

### Classification Scores

### Minimum Distance from /b/, /d/ and /g/ Regression Lines

Using Average Male and Female Coefficients;
All data classified (220 tokens)

| | B | D | G |
|---|---|---|---|
| Correct ID | 85.9% | 70.0% | 65.9% |

Using Male Coefficients;
Female data classified (110 tokens)

| | B | D | G |
|---|---|---|---|
| Correct ID | 75.5% | 69.1% | 62.7% |

Using Female Coefficients;
Male data classified (110 tokens)

| | B | D | G |
|---|---|---|---|
| Correct ID | 90.0% | 48.2% | 56.4% |

spectral shape parameters of the initial burst were used in classification results. Rather, it was the purpose of this study to evaluate the extent to which selected non-burst parameters can correctly classify these consonants.

The 73.9% correct classification score seems at first blush to be much lower than the Kewley-Port (1981) figure of 97% correct identification when the vowel was known (i.e., when consonant classification was done on each vowel separately). However, Kewley-Port's data consists of only one male speaker with five repetitions, so that the classification scores were based on only five separate points per consonant. In this study, the syllables of ten speakers, with two randomizations each were measured, giving twenty separate points to classify for each consonant, and the vowel was not given a priori.

In order to test the effect of inter-subject variability, a discriminant function analysis was done over all ten subjects with two replications for each vowel (i.e., vowel given a priori). The overall jacknife categorization results are given in Table 2.2; on average, the overall correct categorization is 80.6%, with the best results for the vowels /æ/ and /ʌ/ (about 90%). Tables 2.3 to 2.13 show the jacknife categorization results for each vowel in turn.

## TABLE 2.2

### Overall Jacknife Classification Scores for Each Vowel

| Vowel | Overall ID |
|---|---|
| /i/ | 73.3% |
| /I/ | 80.0% |
| /e/ | 81.7% |
| /ɛ/ | 73.3% |
| /æ/ | 88.3% |
| /ʌ/ | 90.0% |
| /ɔ/ | 81.7% |
| /o/ | 85.0% |
| /U/ | 80.0% |
| /u/ | 81.7% |
| /ɜ/ | 71.7% |

## TABLE 2.3

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /i/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 68.4% | 31.8% | 0.0% |
| D | 31.6% | 59.1% | 5.3% |
| G | 0.0% | 9.1% | 94.7% |

## TABLE 2.4

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /I/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 70.0% | 30.0% | 0.0% |
| D | 30.0% | 70.0% | 0.0% |
| G | 0.0% | 0.0% | 100.0% |

## TABLE 2.5

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /e/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 81.0% | 17.6% | 0.0% |
| D | 19.0% | 76.5% | 13.6% |
| G | 0.0% | 5.9% | 86.4% |

## TABLE 2.6

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /ɛ/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 68.2% | 27.8% | 0.0% |
| D | 31.8% | 61.1% | 10.0% |
| G | 0.0% | 11.1% | 90.0% |

## TABLE 2.7

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /æ/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 94.4% | 13.6% | 0.0% |
| D | 5.6% | 81.8% | 5.0% |
| G | 0.0% | 4.6% | 95.0% |

## TABLE 2.8

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /ʌ/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 100.0% | 0.0% | 0.0% |
| D | 0.0% | 81.8% | 11.1% |
| G | 0.0% | 18.2% | 88.9% |

## TABLE 2.9

### Jacknife Classification Scores of Discriminant Function Analysis

### (Vowel /ɔ/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 90.9% | 0.0% | 0.0% |
| D | 0.0% | 78.9% | 26.3% |
| G | 9.1% | 21.1% | 73.7% |

## TABLE 2.10

### Jacknife Classification Scores of Discriminant Function Analysis

### (Vowel /o/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 82.6% | 0.0% | 6.7% |
| D | 0.0% | 86.4% | 6.7% |
| G | 17.4% | 13.6% | 86.6% |

## TABLE 2.11

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /U/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 85.7% | 0.0% | 11.1% |
| D | 0.0% | 81.0% | 16.7% |
| G | 14.3% | 1.9% | 72.2% |

## TABLE 2.12

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /u/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 76.2% | 4.8% | 16.7% |
| D | 0.0% | 90.4% | 5.6% |
| G | 23.8% | 4.8% | 77.7% |

## TABLE 2.13

## Jacknife Classification Scores of Discriminant Function Analysis

## (Vowel /ɚ/)

| Actual Category | Predicted Group Membership (20 tokens) | | |
|---|---|---|---|
| | B | D | G |
| B | 74.0% | 0.0% | 15.8% |
| D | 8.6% | 77.8% | 21.0% |
| G | 17.4% | 22.2% | 63.2% |

Kewley-Port (1980) found that when a discriminant analysis was done for her speaker across all eight vowels in her study (i.e., vowel not identified a priori), a 68% recognition rate was obtained. Thus, the 73.9% correct classification rate obtained by the minimum distance classification rules (from the regression lines of Figure 2.8) seems like only a moderate improvement although, as mentioned above, this study also contends with speaker differences across males and females.

A linear discriminant function analysis was done for all 220 tokens (i.e., both male and female speech with vowels pooled); the results are shown in Table 2.14. When data from the ten subjects are used, the overall correct identification for pooled vowels is 61.8%. Using the same data, a second-order discriminant analysis [4] was done; the results are shown in Table 2.15. This analysis correctly identified 72.4% of the data, which is very close to the overall correct identification rate obtained by the minimum distance classification rules from the regression lines fitted to formant onset vs. steady-state measurements. However, the second-order discriminant function analysis optimizes the criteria of category separation, whereas

---

[4] The analysis reported here actually is a discriminant analysis performed in a two-dimensional linear space with second-order discriminant functions. A true second-order discriminant function would have used second-order discriminant functions in a four dimensional space (SPSS Inc., 1983, p. 604).

## TABLE 2.14

### Classification Scores of Linear Discriminant Function Analysis (Vowels Pooled)

Predicted Group Membership (220 tokens)

| Actual Category | B | D | G |
|---|---|---|---|
| B | 89.5% | 9.1% | 1.4% |
| D | 16.4% | 47.7% | 35.9% |
| G | 15.0% | 25.0% | 60.0% |

Overall Correct ID: 65.76%

## TABLE 2.15

### Classification Scores of Second-Order Discriminant Function Analysis (Vowels Pooled)

Predicted Group Membership (220 tokens)

| Actual Category | B | D | G |
|---|---|---|---|
| B | 84.1% | 10.0% | 5.9% |
| D | 10.5% | 71.4% | 18.2% |
| G | 11.8% | 26.4% | 61.8% |

Overall Correct ID: 72.42%

no such optimization is done using the minimum distance classification algorithm.

As mentioned earlier, it is not the purpose of this study to suggest that formant onsets in relation to the following vowel steady-state are all that is necessary for correct consonant classification. Clearly, burst information is also important. Here we are examining only the extent to which formant information might serve as a cue for stop consonants. In the following chapters, the results of perceptual experiments are presented which show the extent to which the patterns found in this natural data measurement study are important for consonant perception.

# 3. THE PERCEPTION OF FORMANT ONSETS IN RELATION TO F2 STEADY-STATES

This study examines three main questions:

1. To what extent are the measured patterns found in Chapter 2 evident in perception?
2. Can formant measurements and perception be related?
3. Are listeners attending to formant-based information or to spectral shape information?

In order to examine these points, a large-scale perceptual experiment using synthetic speech was designed. Bursts were not synthesized since it was of interest to examine only the contribution of formant parameters.

This study covers the /o/, /U/, /ɛ/ range, in contrast to previous perceptual studies using synthetic speech which focused on vowels at the 'extremes' of the vowel triangle (i.e., /i/, /a/ or /æ/ and /u/). This was done for two reasons:

1. To test whether patterns found by Blumstein and Stevens (1979) for /i/, /a/ and /u/ and by Hoffman (1958) and Harris (1957) for /æ/ would replicate using these vowels.
2. These vowels can be synthesized as a continuum of F2 steady-states, keeping F1 and F3, as well as transition durations constant. Thus, in contrast to Stevens and Blumstein (1979), who altered several vowel parameters simultaneously in order to obtain good tokens of /i/, /a/ and /u/, this study kept all factors constant except

for F2 steady-state of the vowel, and F2 and F3 onsets (synthesizing the consonants). This included time factors which, in other studies, were also manipulated from category to category.

A pilot study was conducted to determine the appropriate F2 range of steady-state values that would cover /o/ - /U/ - /ɛ/, as well as provide for vowel 'boundary' points (i.e., the values for which half of the subjects reported hearing one vowel, while the other half reported hearing another vowel). The boundary point values were used for the 'ambiguous' vowels, [o-U] and [U-ɛ]. Two tests were run, one with Eastern Canadian speakers, and one with Western Canadian speakers. Later experiments were to take place in Western Canada, but it was anticipated that several participating subjects would be from Eastern Canada. Thus it seemed advisable to test in both dialect groups. The pilot experiments are described below.

## 3.1 Pilot Vowel Experiment for Eastern and Western Canadian Dialects

### 3.1.1 Subjects

Ten native Eastern Canadians and nine native Western Canadians were tested. The Easterners were high school students from the Ottawa area and were tested at the University of Ottawa. The Western subjects were graduate students in linguistics originally from Western Canada and were tested at the University of Alberta. None of the subjects had any history of hearing difficulties.

### 3.1.2 Stimuli

Subjects were tested with a synthetic continuum of vowels using an implementation of the Klatt (1980) Speech Synthesizer with a sampling rate of 10 kHz F0 was at 120 Hz, F1 was set at a constant 480 Hz, and F3 was constant at 2350 Hz, while F2 varied in twenty constant log steps which ranged from 850-1800 Hz. (The Eastern Canadian group was also tested with F1 at 400 Hz, but a substantial number of /ɛ/ responses were heard only for the higher F1 value, and so only the results for the higher value are reported.) All stimuli were 200 msec long. Stimuli for the Eastern Canadian group were constructed on the PDP 11/34 computer at the Phonetics Laboratory at Brown University and taped for presentation in Ottawa, while the stimuli for the Western Canadian group were constructed on the PDP-12 Computer at the University of Alberta and were presented on-line. Ten randomizations were presented to subjects. Each stimulus was presented twice before proceeding to the next stimulus.

### 3.1.3 Procedure

Taped versions of the stimuli were presented to the Eastern Canadians using a TEAC A-7030 tape recorder. The Eastern Canadians were required to write down their vowel responses as either "o" for /o/, "u" for /U/ and "e" for

/ɛ/. Appropriate English words were given as examples of each vowel sound. The Western Canadians responded by pressing appropriately marked switches on a switch box, and their responses were automatically recorded.

### 3.1.4 Results and Analysis

Responses of the Eastern Canadian group were hand-tallied and scored; the Western Canadian responses were computer scored. The results are shown in Figure 3.1. The [o-U] boundary for both Eastern and Western Canadian are nearly identical (at approximately 1050 Hz). However, the [U-ɛ] boundary for Western Canadians is higher than that of Eastern Canadians. While Easterners change from /U/ to /ɛ/ at approximately 1450 Hz, the Westerners' [U-ɛ] boundary is at approximately 1625 Hz. The Western boundaries can be compared with the vowel measurements of average formant values of Western Canadian English provided by Assmann (1979). The average male /o/ has F2=980 Hz, the average male /U/ has F2=1176 Hz, and the average male /ɛ/ has F2=1793 Hz. The results are also consistent with Assmann's (1979) finding that /U/'s cover a wide range of F2 values.

Unfortunately, to the author's knowledge, no measurements of Eastern Canadian English have been previously reported. However, from the perception data presented here, it would not be surprising to find that average Eastern Canadian /ɛ/'s have lower average F2's than Western /ɛ/'s or, alternatively, that the Eastern /U/

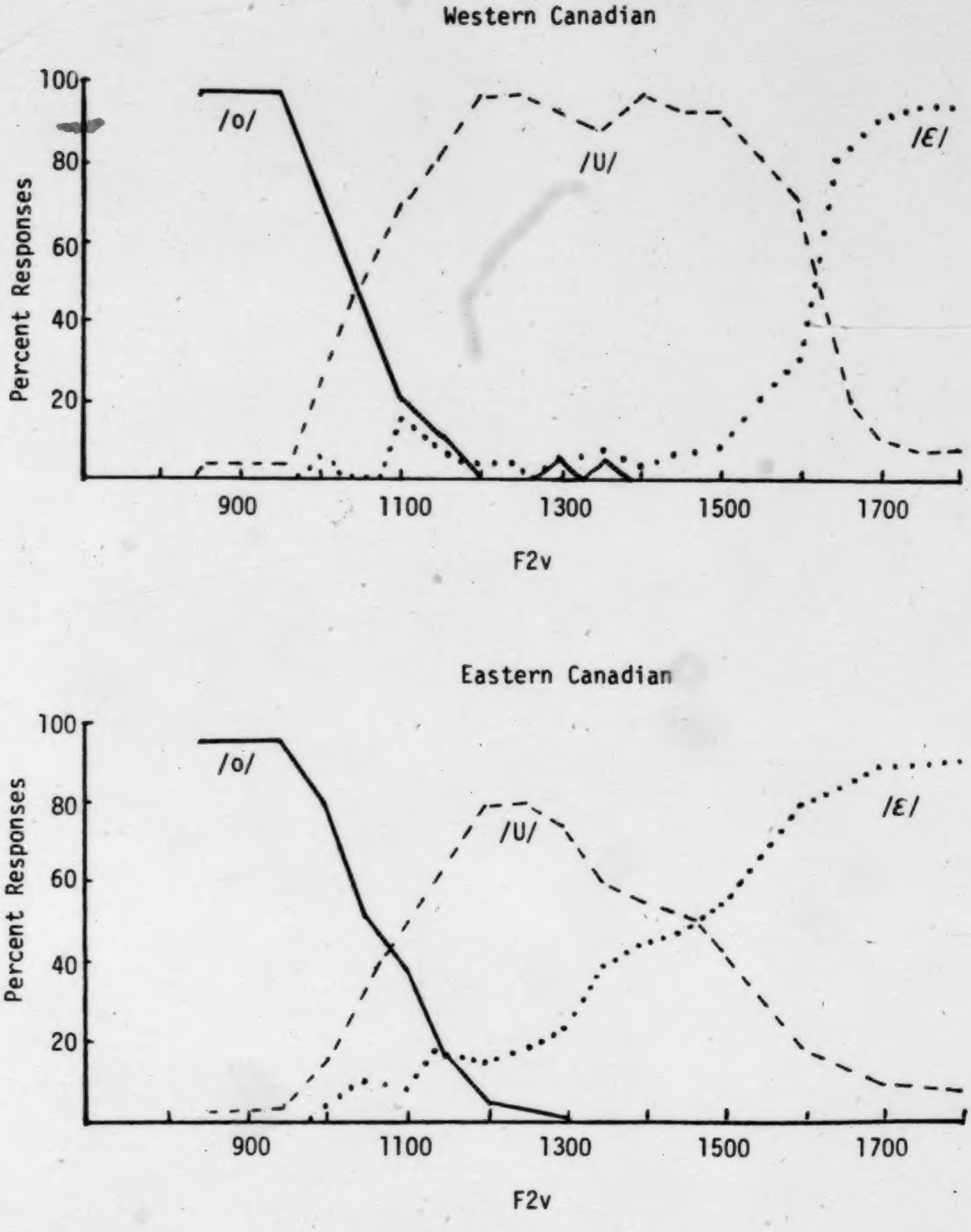


FIGURE 3.1 Vowel Categorization of Western Canadian and Eastern Canadian Listeners

category covers a smaller F2 range (cf. Willis, 1971 on the comparison of dialect or language groups with vowel perception of synthetic stimuli.)

From this pilot test, appropriate formant values for /o/, /U/ and /ɛ/, as well as two 'ambiguous' vowels, [o-U] and [U-ɛ] could be chosen for the large scale synthetic experiment presented below for both Eastern and Western Canadian speakers.

## 3.2 MAIN PERCEPTION EXPERIMENT: THE ROLE OF F2v

### 3.2.1 Subjects

Twenty-two university students were tested. All subjects had taken at least one undergraduate linguistics course which included some practical training in phonetic transcription. No subject had a history of speech or hearing difficulties, and all were native speakers of Canadian English.

### 3.2.2 Stimuli

Five synthetic vowels were synthesized using an implementation of Klatt's (1980) software synthesizer based on the 'Pilot Vowel Experiment'. Stimuli were constructed on a PDP-12 computer with a sampling rate of 10 kHz. F0 was set at 120 Hz, F1=480 Hz and F3=2350 Hz for all vowels. The F2 values selected from the pilot test were as follows: F2=900 Hz, for /o/; F2=1050 Hz for ambiguous [o-U]; F2=1200 Hz for

/U/; F2=1450 Hz for ambiguous [U-ε]; and F2=1650 Hz for /ε/. All vowels were 200 msec in duration.

To synthesize the consonants of the CV syllables, F1, F2 and F3 formant trajectories were appended to each vowel. In all cases, the F1 onset was set at 180 Hz, with a sharp 20 msec rise to the F1 of the vowel. This was necessary to produce a 'stop' quality to the stimuli in the absence of bursts (see Blumstein and Stevens, 1980). A continuum of F2 and F3 onsets was synthesized for each vowel in 100 Hz steps. All combinations of F2 onsets (F2i) crossed with F3 onsets (F3i) were synthesized, except for those values where F3i would be lower or equal to F2i. Onsets were chosen, based on pilot experiments, to ensure categorizations of all three stop consonants /b/, /d/ and /g/. F2i and F3i reached the steady-state targets in 30 msec. Transition times for F2 and F3 were identical to those in Blumstein and Stevens (1980) synthetic stimuli.

The formant onset values were as follows:

1. For /o/, F2i ranged from 700 Hz to 1600 Hz, and F3i ranged from 1500 Hz to 2400 Hz.
2. For ambiguous [o-U], F2i ranged from 900 Hz to 1700 Hz, and F3i ranged from 1600 Hz to 2400 Hz.
3. For /U/, F2i ranged from 900 Hz to 1700 Hz, and F3i ranged from 1600 Hz to 2400 Hz.
4. For ambiguous [U-ε], F2i ranged from 1200 Hz to 1800 Hz, and F3i ranged from 1800 Hz to 2400 Hz.
5. For /ε/, F2i ranged from 1400 Hz to 2000 Hz, and F3i

ranged from 2100 Hz to 2600 Hz.

A schematic diagram of the stimuli is shown in Figure 3.2. Stimuli were randomized and taped for presentation. Two randomizations were done, and an equal number of subjects were randomly assigned to listen to each of the stimulus sets. Each stimulus was presented twice.

### 3.2.3 Procedure

Subjects were instructed to write down both a consonant and vowel response for each stimulus. The consonant response was limited to one of /b/, /d/ or /g/, while the vowel response was a free choice. However, subjects were told that they might hear the vowels /o/, /U/ and /ɛ/, but if they felt that another vowel was being presented, they were free to respond with that vowel. Since all subjects had some phonetic training, IPA symbols were used.

### 3.2.4 Results and Analysis

Figure 3.3 shows the consonant categorizations pooled over subjects and vowels. The categorization data shows that F2i, F3i and F2v all affect consonant categorization. The main findings are as follows:

1. For low F2v's (F2v=900 Hz or 1050 Hz), low F2i's were heard as /g/, while high F2i's generally were heard as /d/'s. However, a high F2i combined with a very low F3i produced more /g/ responses. A rather surprising result was that /b/'s were generally not reliably heard at any

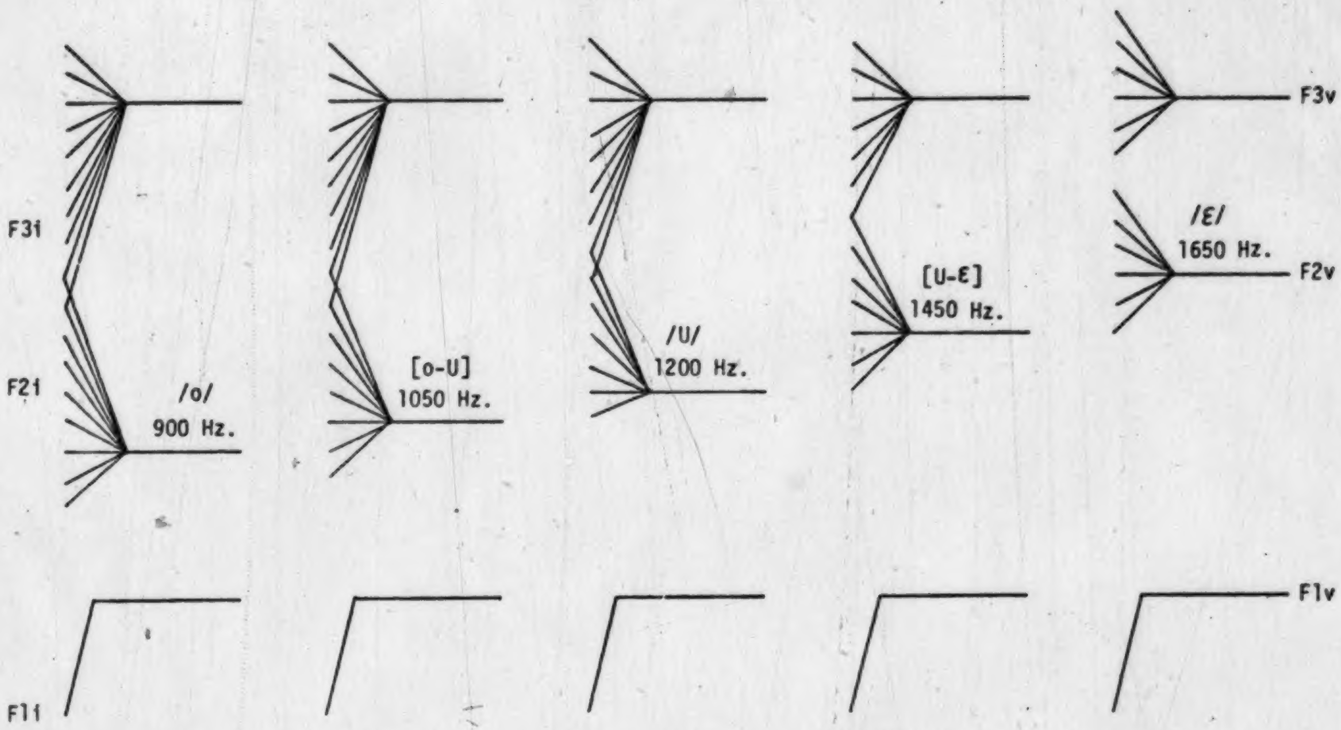

F3i
F2i
F1i
/o/
900 Hz.
[o-U]
1050 Hz.
/U/
1200 Hz.
[U-ɛ]
1450 Hz.
/ɛ/
1650 Hz.
F3v
F2v
F1v

FIGURE 3.2 Schematic Diagram of Stimuli; F2v Experiment

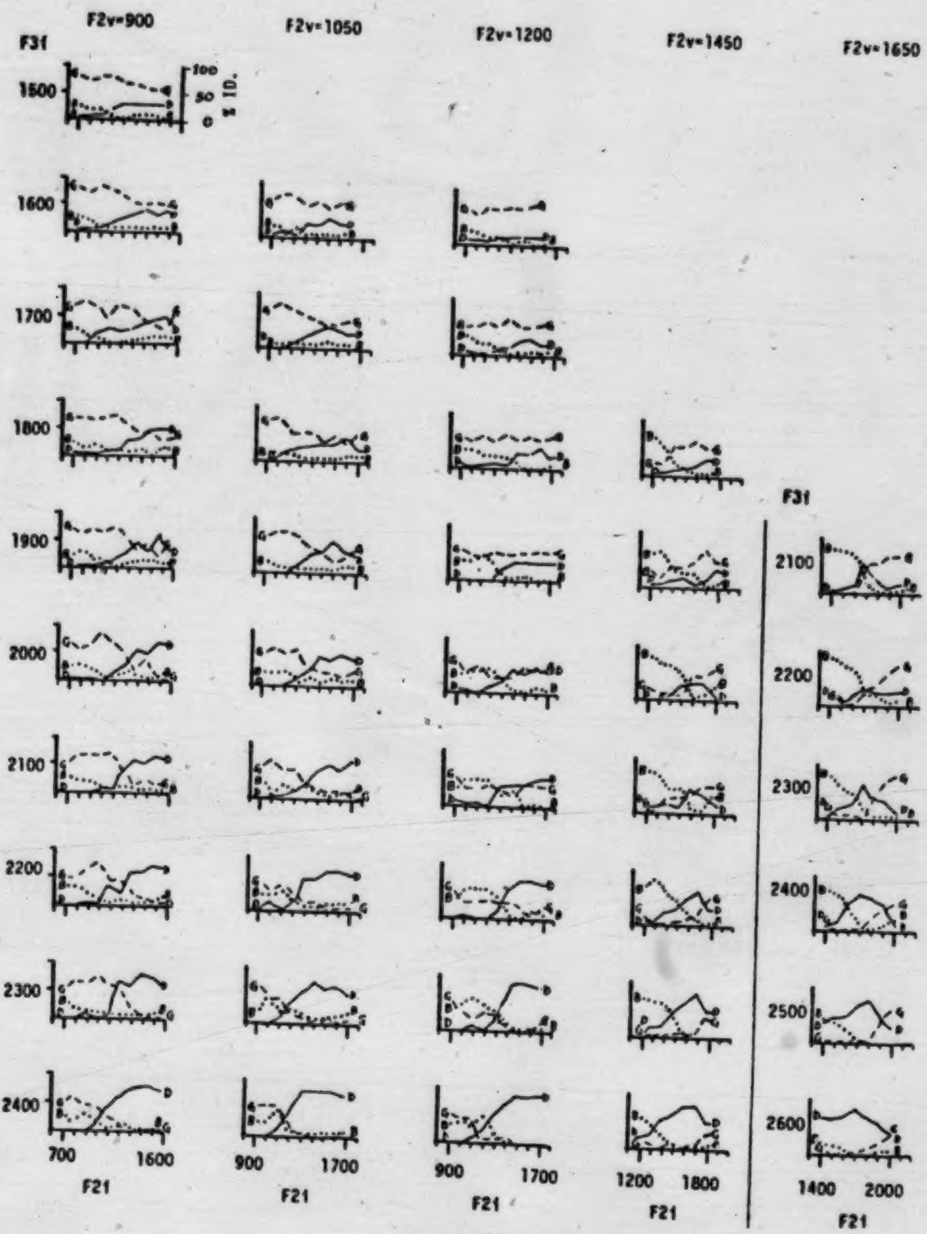


FIGURE 3.3 Consonant Categorizations; F2v Experiment
(22 Subjects Pooled)

of the F2i or F3i values.

2. For F2v=1200 Hz, low F2i's were heard as either /g/ or /b/. The F3i's affected listeners' responses; for low F2i's, low F3i's were heard as /g/, while high F3i's were heard as /b/. For high F2i's, low F3i's cued /g/, while high F3i's cued /d/. Also, /b/ responses never greatly exceeded /g/ responses.
3. For F2v=1450 Hz, low F2i's were heard as /b/, while higher F2i's were heard as /g/ or /d/. Again, F3i was an important factor in stimuli which had high F2i. For these stimuli, low F3i's were heard as /g/, while high F3i's were heard as /d/.
4. For F2v=1650 Hz, low F2i's were heard mainly as /b/, but could also be heard as /d/ if F3i was high. Intermediate values of F2i produced mainly /d/ responses, though low F3i values could shift majority responses to /g/. High F2i values were heard as /g/.

The results obtained for F2v=1650 Hz are similar to results obtained by Hoffman (1958) for the vowel /æ/ (where the F2v was approximately 1650 Hz). In the present experiment, however, a radically different synthesis technique was used (cascade formant synthesis as opposed to the harmonic synthesis done by the Pattern Playback in Hoffman, 1958).

Do the results of this study reflect patterns in production data? Both F2i and F3i contribute to place of articulation categorization. The measurement study of

Chapter 2 also showed that F2i and F3i were different, depending on place of articulation. For low F2v values, low F2i's are heard as /b/ or /g/, while high F2i's are heard as /d/. In natural data measurement (see Figures 2.8 and 2.9) low F2v's have low F2i's for /b/ and /g/, and higher F2i's for /d/. Thus, perception results correspond to some degree with measurement of natural data. For higher F2v's, measurement showed that /g/'s had high F2i's, while /d/ had lower F2i's. Again, the perception study reflects this pattern, since high F2i's were heard as /g/'s while low F2i's were heard as /d/'s when F2v was high (for F2v=1650 Hz). Measurement also showed that /g/'s have lower F3i's than /d/'s, and /b/'s have lower F3i's than /d/'s at the F3v value which was used here (F3v=2350 Hz), a pattern that is also reflected in this perception study.

Results generally concur with previous studies (Blumstein and Stevens, 1980, Cooper et al., 1952) though, in this study, majority /b/ responses were not obtained for low F2v with low F2i values, as expected. Blumstein and Stevens (1980), for example, obtained a weak majority of /b/ responses (over 60%) for F2v=1100 Hz (vowel /u/) with a 'synthetic' burst present, and a majority /b/ response of close to 80% for transition only stimuli (i.e., without a synthetic burst). However, it should be noted that Blumstein and Stevens (1980) also manipulated other aspects of the signal, such as transition duration, which may also be affecting listeners' responses. In addition, their lowest

F2i appended onto this F2v value (1100 Hz) was slightly lower than in this experiment (800 Hz in their study as opposed to 900 Hz in this study). In this experiment, perhaps lower F2i's could have been included for the vowel sets having lower F2v's, and then more /b/ responses could have possibly emerged. The results of Cooper et al. (1952) also showed majority /b/ responses for low F2i onsets in front of vowels /o/ and /u/ (which have low F2v's), though /g/ responses were also frequent in their study.

However, the main emphasis of this study was the 'crossover' of /d/ and /g/ responses, depending upon changes in F2v. This experiment showed that perception follows measured data in this regard: high F2i's cued /d/, when followed by low F2v's, but cued /g/ when followed by high F2v's. Furthermore, this change in perception occurred approximately at the F2v 'cutoff' point that was found in measurement (i.e., cutoff from male measurement data in Figure 2.8 was F2v=1550 Hz; change in perception ocurred for stimuli with F2v=1650 Hz.)

With regard to the '/d/-/g/ crossover', the results of this perception study corroborates Cooper et al.'s (1952) results. They found that onsets before vowels having low F2v's were perceived as /d/ when the onsets were low and as /g/ when the onsets were high, while onsets before vowels having high F2v's were perceived as /g/ when the onsets were low and /d/ when the onsets were high. In this experiment, *using a completely different synthesis technique, different*

*F2v values, and having F3 present*, subjects' responses were quite parallel to Cooper et al.'s results: For lower F2v's, low onset values were heard as /g/ and high onset values were heard as /d/, while for higher F2v's, low onset values were heard as /d/'s and high onset values were heard as /g/'s.

### 3.2.4.1 Analysis of Acoustic Features

What features were being attended to by listeners? Certainly, F2i, F3i and F2v influenced listeners' stop consonant categorization. Could responses be interpreted as if subjects were somehow 'tracking' the regression lines of formant onsets in relation to vowel steady-states? Or could subjects' responses be interpreted as a response to gross spectral shape at stimulus onset (Stevens and Blumstein 1980), a relative change in gross spectral shape (Lahiri and Blumstein, 1982), or perhaps a temporal track of spectral shapes (Kewley-Port, 1980)?

The fact that in this study the '/d/-/g/ crossover' (i.e., change in /d/ and /g/ responses as a function of vowel steady-state) was similar to the study done by Cooper et al. (1952) suggests that it is formant frequency relationships that are important, since Cooper et al.'s synthetic stimuli were constructed without F3 and thus the onset spectral shapes could not have been highly diffuse. However, exact onset spectral shapes of their synthetic stimuli can not be reconstructed from these early experiments.

An examination of the roles of formant information and spectral shape was done for all stimuli. It is important to note that, in cascade synthesis, the formant onsets and onset spectral shape are not entirely independent. If, for example, a stimulus had F2 and F3 formant onsets that were close together, its onset spectral shape would be more 'compact'. High formant onsets would tend to produce rising onset spectra and low formant onsets would tend to produce falling spectra. Nevertheless, formant and spectral shape information are treated as separate cues since spectral and formant information can be made to provide conflicting cues when parallel synthesis is used (Blumstein et al., 1982; Walley and Carrel, 1983).

In order to test the role of formant information, predictions of consonant categorizations were made based on the formant measurement study of Chapter 2. The role of spectral shape was investigated by fitting spectral shape templates to the synthetic stimuli. LPC 3-D running plots (similar to those used by Kewley-Port, 1980) of the first 40 msec of each stimuli were done using programs written by Dr. T. M. Nearey. The window length was set to 16 msec. Eight spectral shapes were plotted for each stimulus, taken 5 msec apart. Following Stevens and Blumstein (1979), well-identified stimuli were compared with non-well-identified stimuli in order to see which features were important for listeners. If 80% or more of subjects identified a stimulus as belonging to a particular consonant category, that stimulus was considered to be well identified, or a 'good' token. If a stimulus was identified by approximately half of the subjects as one consonant, and by the other half as another consonant, that stimulus was considered to be poorly identified, or an 'ambiguous' token. Some stimuli were ambiguous between all three categories.

Figures 3.4 through 3.8 show examples of stimuli that were both well and poorly identified for each F2v value. Figure 3.4 shows three 'good' tokens of /go/, each of which obtained 91% of the number of subjects' responses, a 'good' /do/ which obtained 95% responses and an example of an 'ambiguous' stimulus between /d/ and /g/ (denoted as [d-g]) which obtained 50% /d/ responses and 41% /g/ responses. Three 'good' /g/ tokens were chosen since they had rather different formant onset values and it was of interest to compare their onset spectral shapes.

Figure 3.5 shows good tokens of [go/gU] (which obtained 91% responses), and [do/dU] (95% responses), an 'ambiguous' [d-g] (45% /d/ and /g/ responses) and an 'ambiguous', [b-d-g] (36% /d/ and /g/ responses, 28% /b/ responses). Note that for 'ambiguous' stimuli in this F2v set, both the consonant and vowel are ambiguous.

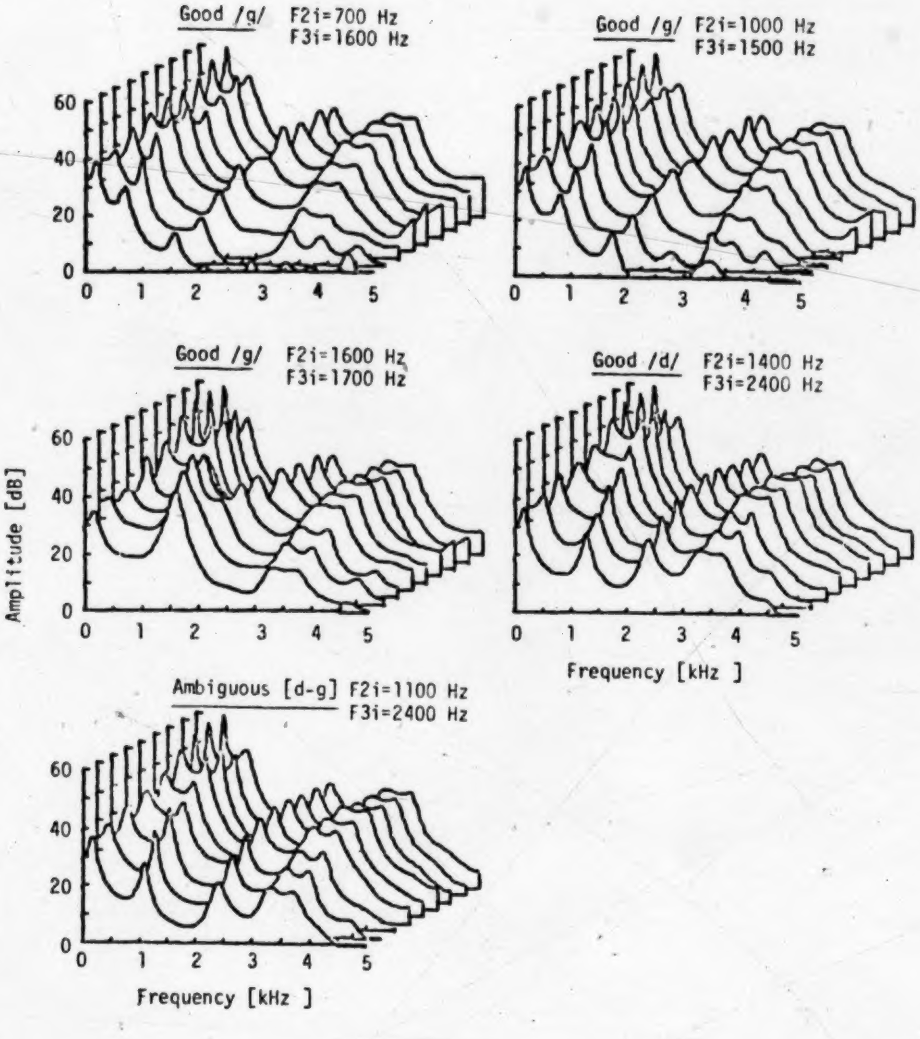


FIGURE 3.4 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /o/ (F2v=900 Hz )

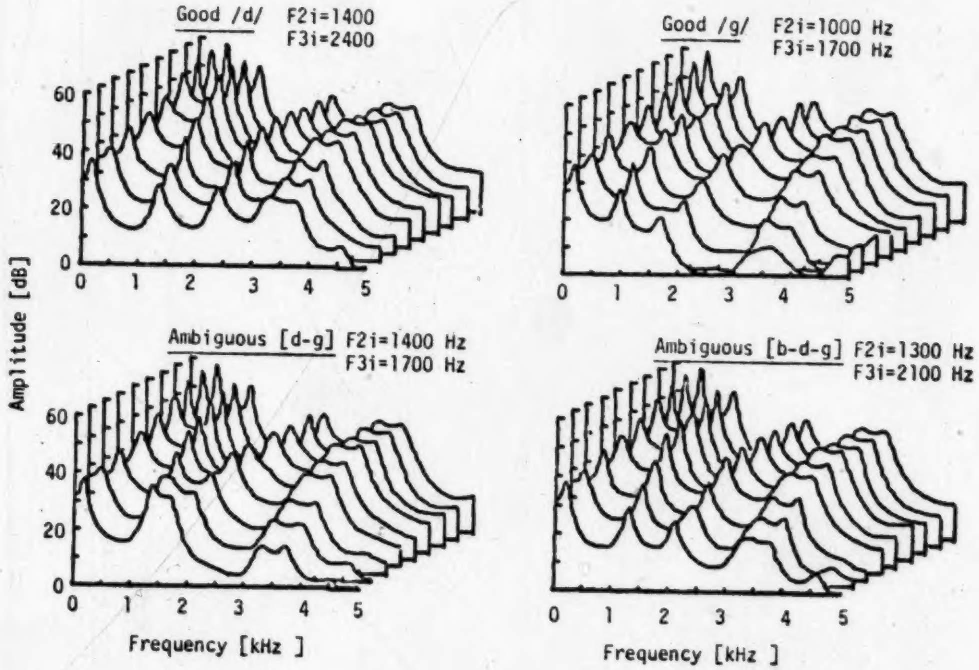


FIGURE 3.5 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel [o-U] (F2v=1050 Hz )

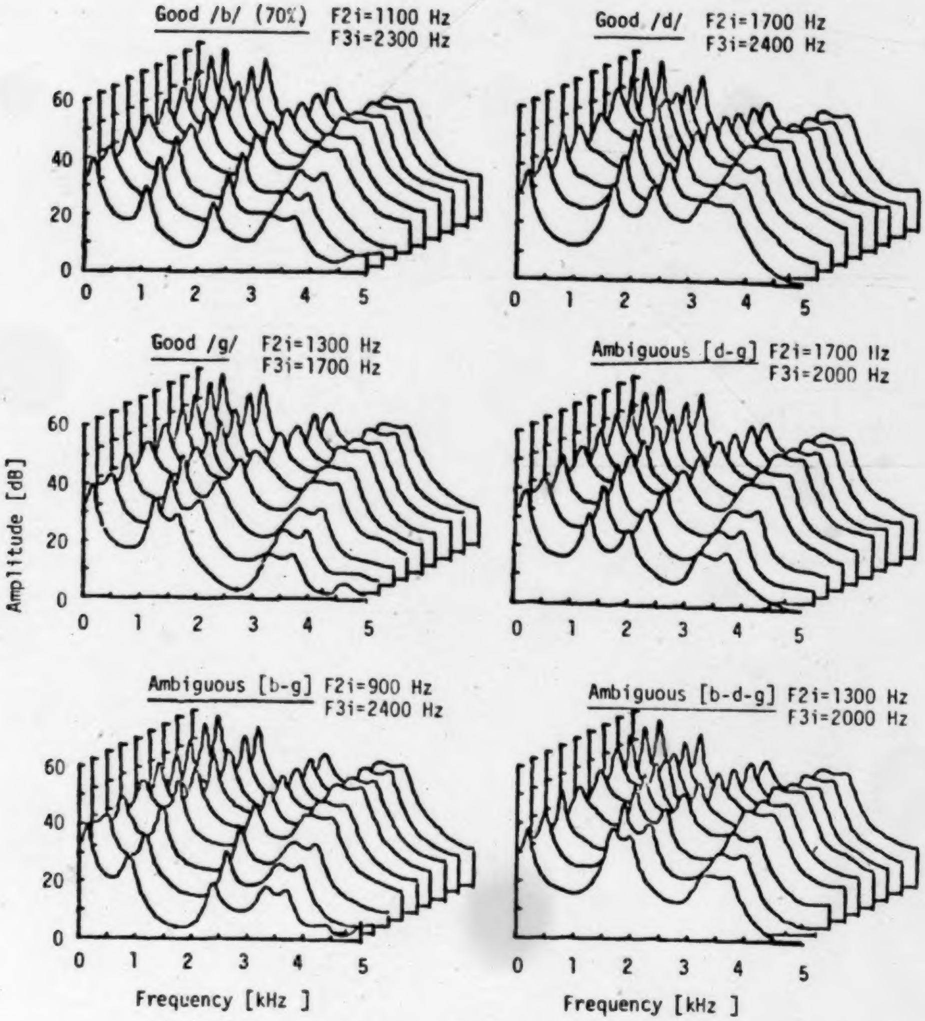
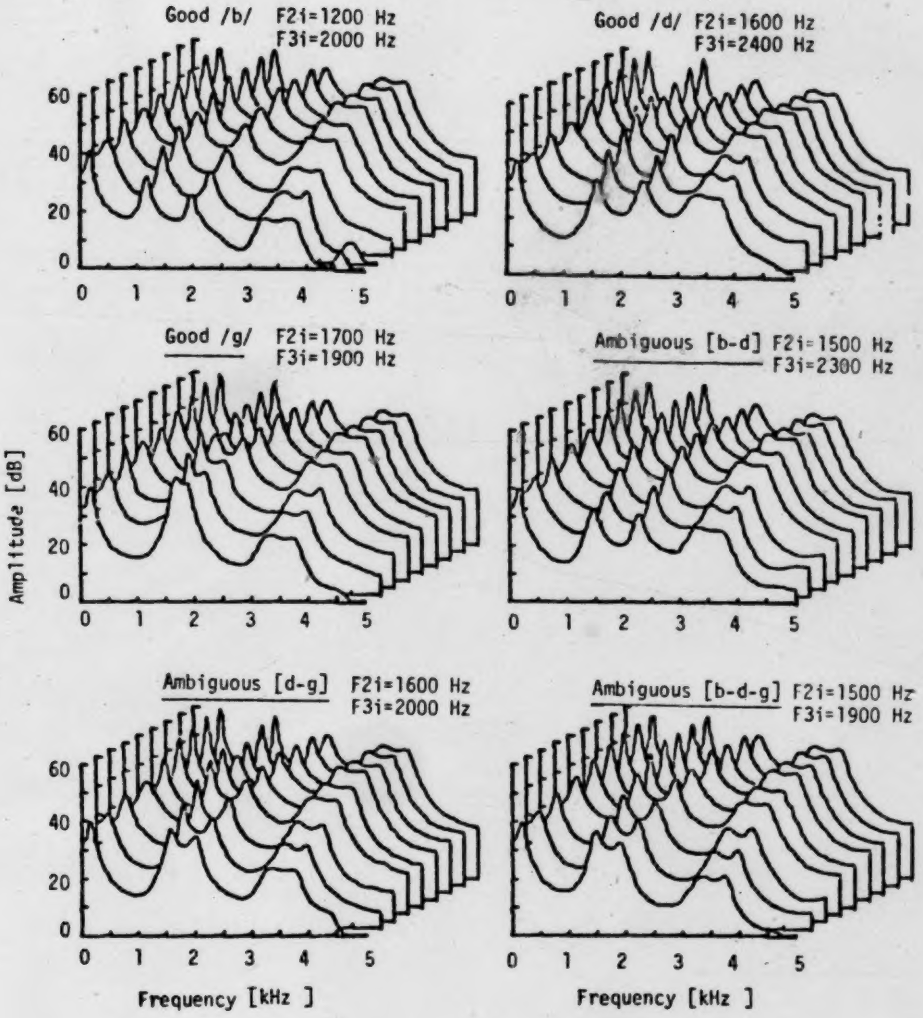


FIGURE 3.6 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /U/ (F2v=1200 Hz )



Good /b/ F2i=1200 Hz F3i=2000 Hz

Good /d/ F2i=1600 Hz F3i=2400 Hz

Good /g/ F2i=1700 Hz F3i=1900 Hz

Ambiguous [b-d] F2i=1500 Hz F3i=2300 Hz

Ambiguous [d-g] F2i=1600 Hz F3i=2000 Hz

Ambiguous [b-d-g] F2i=1500 Hz F3i=1900 Hz

FIGURE 3.7 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel [U-ɛ] (F2v=1450 Hz )

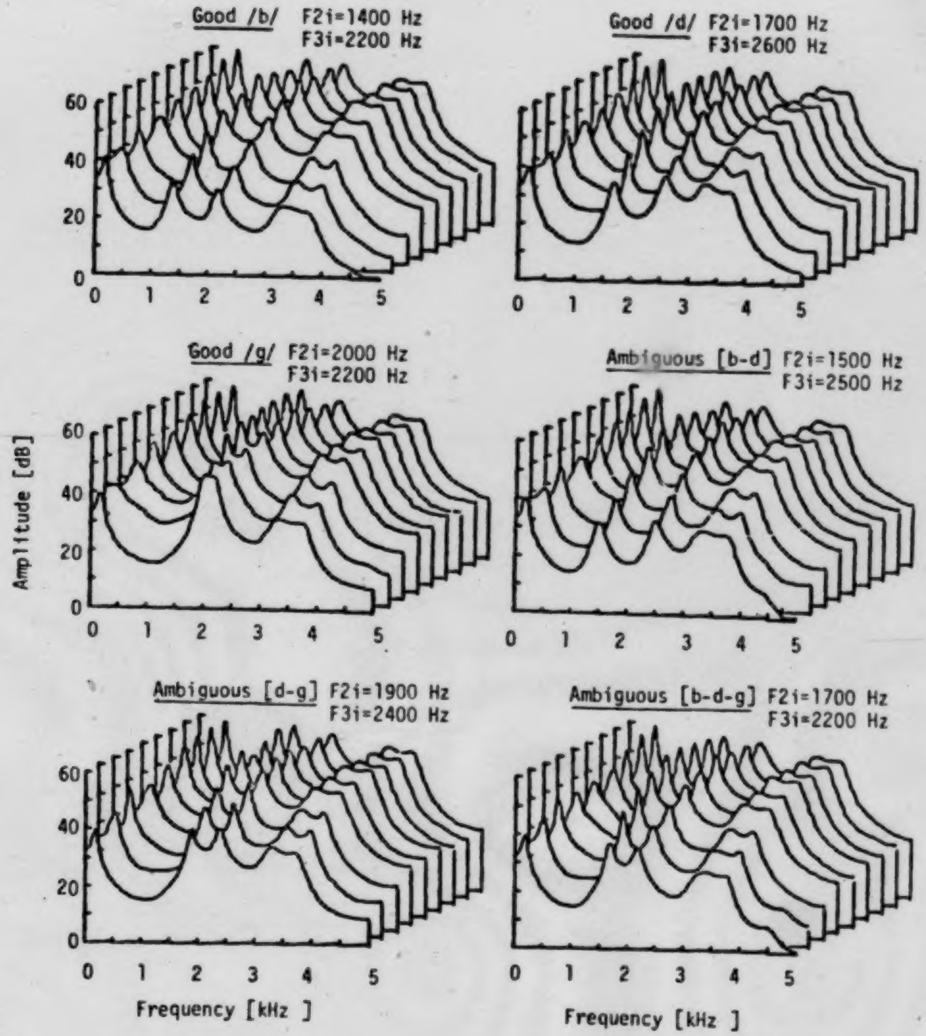


FIGURE 3.8 Running Spectra of Well Identified and Poorly Identified Stimuli; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz)

Figure 3.6 shows 'good' tokens of /gU/ (82% responses) and /dU/ (100% responses), and 'ambiguous' stimuli [gU/bU] (55% /g/ responses, 45% /b/ responses), [dU/gU] (50% /d/ responses and 50% /g/ responses), and [bU/gU/dU] (41% /b/ responses, 32% /g/ responses and 27% /d/ responses). Also shown is a token of /bU/ that, while it did not meet the 80% criterion, nevertheless received 70% of the subjects' responses.

Figure 3.7 shows 'good' tokens of [bU/bɛ] (91% responses), [dU/dɛ] (91% responses) and [gU/gɛ] (82% responses), as well as 'ambiguous' [b/d/g] (41% /g/ responses, 32% /b/ responses and 27% /d/ responses), 'ambiguous'[d/g] (50% /g/ responses, 41% /d/ responses) and [b/d] (55% /b/ responses and 36% /d/ responses). For 'ambiguous' stimuli in this vowel set, both the consonant and vowel are ambiguous.

Figure 3.8 shows 'good' tokens of /bɛ/ (100% responses), /dɛ/ (100% responses) and /gɛ/ (86% responses), as well as 'ambiguous' [bɛ/dɛ] (55% /d/ responses, 45% /b/ responses), 'ambiguous' [dɛ/gɛ] 55% /d/ responses, 45% /g/ responses) and 'ambiguous' [bɛ/dɛ/gɛ] (36% /d/ and /g/ responses, 28% /b/ responses).

Inspection of these plots shows that formant onset values, as well as onset spectral shapes, both play an important role in the identification of these stimuli. The role of formant onsets in relation to the vowel steady-state will first be discussed, followed by an evaluation of the gross spectral shape at stimulus onset.

### 3.2.5 The Use of Formant Information

From the regression lines of Figure 2.8 (male measurement data), predicted values of F2i and F3i for *b*, *d*, and *g* were calculated for each F2v and F3v. (Since F3v was kept constant, the predicted F3i remained constant for each consonant category). The predicted values are shown in Table 3.1.

Looking at those stimuli which were well and poorly identified in Figures 3.4-3.8, it can be seen that these predictions are good only for high F2v's (i.e., F2v=1650 Hz). Tables 3.2-3.6 (for F2v=900 Hz, 1050 Hz, 1200 Hz, 1450 Hz and 1650 Hz, respectively) show the observed category response and the predicted category response from the measurement data for each of the 'good' and 'ambiguous' stimuli. The regression line which had the minimum deviation from the stimulus value was the 'predicted category'. Deviation of the stimulus onset values from the regression lines were calculated in terms of Hz difference and percentages. The F2i and F3i values of each stimulus are also given. The details of this examination are presented below.

F2v=900 Hz (Vowel /o/)

For F2v=900 Hz we would expect that a stimulus with F2i

## TABLE 3.1

## Predicted Values of F2i and F3i From Male Formant Measurements; F2v Experiment

| F2v | F3v | | F2i | F3i |
|---|---|---|---|---|
| 900 | 2350 | b | 935 | 2346 |
| | | d | 1448 | 2525 |
| | | g | 1083 | 2491 |
| 1050 | 2350 | b | 1062 | 2346 |
| | | d | 1518 | 2525 |
| | | g | 1240 | 2491 |
| 1200 | 2350 | b | 1190 | 2346 |
| | | d | 1590 | 2525 |
| | | g | 1398 | 2491 |
| 1450 | 2350 | b | 1402 | 2346 |
| | | d | 1706 | 2525 |
| | | g | 1660 | 2491 |
| 1650 | 2350 | b | 1529 | 2346 |
| | | d | 1780 | 2525 |
| | | g | 1818 | 2491 |

## TABLE 3.2

### Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /o/ (F2v=900 Hz )

| Obtained Category | Predicted Category | | | | Deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F2i | F3i | F2i & F3i | | F2i | | F3i | | F2i & F3i |
| | | | | | Hz | % | Hz | % | Hz |
| Good /g/ | /b/ | /b/ | /b/ | b | -235 | 25.1 | -747 | 32.0 | 982 |
| F2i=700 | | | | d | -748 | 51.7 | -925 | 37.0 | 1673 |
| F3i=1600 | | | | g | -383 | 35.4 | -891 | 36.0 | 1274 |
| Good /g/ | /b/ | /b/ | /b/ | b | 65 | 6.95 | -847 | 36.0 | 912 |
| F2i=1000 | | | | d | -448 | 30.9 | -1025 | 40.6 | 1473 |
| F3i=1500 | | | | g | -83 | 7.6 | -991 | 39.8 | 1074 |
| Good /g/ | /d/ | /b/ | /d/ | b | 665 | 71.1 | -646 | 27.5 | 1311 |
| F2i=1600 | | | | d | 152 | 10.5 | -825 | 32.7 | 977 |
| F3i=1700 | | | | g | 517 | 47.7 | -791 | 31.8 | 1308 |
| Good /d/ | /d/ | /b/ | /d/ | b | 465 | 49.7 | 54 | 2.3 | 519 |
| F2i=1400 | | | | d | -48 | 3.32 | -125 | 4.95 | 173 |
| F3i=2400 | | | | g | 317 | 29.3 | -91 | 3.7 | 408 |
| Ambiguous [d-g] | /g/ | /b/ | /g/ | b | 165 | 17.7 | 54 | 2.3 | 219 |
| F2i=1100 | | | | d | -348 | 24.0 | -125 | 4.95 | 473 |
| F3i=2400 | | | | g | 17 | 1.57 | -91 | 3.7 | 108 |

## TABLE 3.3

### Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel [o-U] (F2v=1050 Hz )

| Obtained Category | Predicted Category | | | | Deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F2i | F3i | F2i & F3i | | F2i | | F3i | | F2i & F3i |
| | | | | | Hz | % | Hz | % | Hz |
| Good /d/<br>F2i=1400<br>F3i=2400 | /d/ | /b/ | /d/ | b | 338 | 31.8 | 54 | 2.3 | 392 |
| | | | | d | -118 | 7.8 | -125 | 4.95 | 243 |
| | | | | g | 160 | 12.9 | -91 | 3.7 | 251 |
| Good /g/<br>F2i=1000<br>F3i=1700 | /b/ | /b/ | /b/ | b | -62 | 5.8 | -646 | 27.5 | 708 |
| | | | | d | -518 | 34.1 | -825 | 32.7 | 1343 |
| | | | | g | -240 | 19.4 | -791 | 31.8 | 1031 |
| Ambiguous<br>[d-g]<br>F2i=1400<br>F3i=1700 | /d/ | /b/ | /d/ | b | 338 | 31.8 | -646 | 27.5 | 984 |
| | | | | d | -118 | 7.8 | -825 | 32.7 | 943 |
| | | | | g | 160 | 12.9 | -791 | 31.8 | 951 |
| Ambiguous<br>[d-b-g]<br>F2i=1300<br>F3i=2100 | /g/ | /b/ | /g/ | b | 238 | 22.4 | -246 | 10.5 | 484 |
| | | | | d | -218 | 14.7 | -425 | 16.8 | 643 |
| | | | | g | 60 | 4.9 | -391 | 15.7 | 451 |

## TABLE 3.4

## Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /U/ (F2v=1200 Hz )

| Obtained Category | Predicted Category | | | | Deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F2i | F3i | F2i & F3i | | F2i | | F3i | | F2i & F3i |
| | | | | | Hz | % | Hz | % | Hz |
| Good /b/<br>F2i=1100<br>F3i=2300 | /b/ | /b/ | /b/ | b | -190 | 15.9 | -46 | 1.9 | 236 |
| | | | | d | -490 | 30.8 | -225 | 8.9 | 715 |
| | | | | g | -298 | 21.3 | -192 | 7.7 | 490 |
| Good /d/<br>F2i=1700<br>F3i=2400 | /d/ | /b/ | /d/ | b | 510 | 42.9 | 54 | 2.3 | 564 |
| | | | | d | 110 | 6.9 | -125 | 4.95 | 235 |
| | | | | g | 302 | 21.6 | -91 | 3.7 | 393 |
| Good /g/<br>F2i=1300<br>F3i=1700 | /b/ | /b/ | /b/ | b | 110 | 9.3 | -646 | 27.5 | 756 |
| | | | | d | -290 | 18.2 | -825 | 32.7 | 1115 |
| | | | | g | -98 | 7.0 | -791 | 31.8 | 889 |
| Ambiguous<br>[d-g]<br>F2i=1700<br>F3i=2000 | /d/ | /b/ | /d/ | b | 510 | 42.9 | -346 | 14.7 | 856 |
| | | | | d | 110 | 6.9 | -525 | 20.8 | 635 |
| | | | | g | 302 | 21.6 | -492 | 19.7 | 794 |
| Ambiguous<br>[b-g]<br>F2i=900<br>F3i=2400 | /b/ | /b/ | /b/ | b | -290 | 24.4 | 54 | 2.3 | 344 |
| | | | | d | -690 | 43.4 | -125 | 4.95 | 815 |
| | | | | g | -498 | 35.6 | -91 | 3.7 | 589 |
| Ambiguous<br>[b-d-g]<br>F2i=1300<br>F3i=2000 | /b/ | /b/ | /b/ | b | 110 | 9.3 | -346 | 14.7 | 456 |
| | | | | d | -290 | 18.2 | -525 | 20.8 | 815 |
| | | | | g | -98 | 7.0 | -492 | 19.7 | 590 |

## TABLE 3.5

## Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel [U-ɛ] (F2v=1450 Hz )

| Obtained Category | Predicted Category | | | | Deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F2i | F3i | F2i & F3i | | F2i | | F3i | | F2i & F3i |
| | | | | | Hz | % | Hz | % | Hz |
| Good /b/<br>F2i=1200<br>F3i=2000 | /b/ | /b/ | /b/ | b | -202 | 14.4 | -346 | 14.7 | 548 |
| | | | | d | -506 | 29.7 | -525 | 20.8 | 1031 |
| | | | | g | -460 | 27.7 | -492 | 19.7 | 952 |
| Good /d/<br>F2i=1600<br>F3i=2400 | /g/ | /b/ | /g/ | b | 198 | 14.1 | 54 | 2.3 | 252 |
| | | | | d | -106 | 6.2 | -125 | 4.95 | 231 |
| | | | | g | -60 | 3.6 | -91 | 3.7 | 151 |
| Good /g/<br>F2i=1700<br>F3i=1900 | /d/ | /b/ | [d-g] | b | 298 | 21.3 | -446 | 19.0 | 744 |
| | | | | d | -6 | 0.4 | -625 | 24.8 | 631 |
| | | | | g | 40 | 2.4 | -591 | 23.7 | 631 |
| Ambiguous<br>[b-d]<br>F2i=1500<br>F3i=2300 | /b/ | /b/ | /b/ | b | 98 | 7.0 | -46 | 1.9 | 144 |
| | | | | d | -206 | 12.1 | -225 | 8.9 | 431 |
| | | | | g | -160 | 9.6 | -192 | 7.7 | 352 |
| Ambiguous<br>[d-g]<br>F2i=1600<br>F3i=2000 | /g/ | /b/ | /b/ | b | 198 | 14.1 | -346 | 14.7 | 544 |
| | | | | d | -106 | 6.2 | -525 | 20.8 | 631 |
| | | | | g | -60 | 3.6 | -492 | 19.7 | 552 |
| Ambiguous<br>[b-d-g]<br>F2i=1500<br>F3i=1900 | /b/ | /b/ | /b/ | b | 98 | 7.0 | -446 | 19.0 | 544 |
| | | | | d | -206 | 12.1 | -625 | 24.8 | 831 |
| | | | | g | -160 | 9.6 | -591 | 23.7 | 751 |

## TABLE 3.6

### Obtained vs. Predicted Response Categories From Male Formant Measurements; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz )

| Obtained Category | Predicted Category | | | | Deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F2i | F3i | F2i & F3i | | F2i | | F3i | | F2i & F3i |
| | | | | | Hz | % | Hz | % | Hz |
| Good /b/ | /b/ | /b/ | /b/ | b | -129 | 8.4 | -146 | 6.2 | 275 |
| F2i=1400 | | | | d | -380 | 21.3 | -325 | 12.9 | 705 |
| F3i=2200 | | | | g | -418 | 23.0 | -291 | 11.7 | 709 |
| Good /d/ | /d/ | /d/ | /d/ | b | 171 | 11.2 | 254 | 10.8 | 425 |
| F2i=1700 | | | | d | -80 | 4.5 | 75 | 3.0 | 155 |
| F3i=2600 | | | | g | -118 | 6.5 | 109 | 4.4 | 227 |
| Good /g/ | /g/ | /b/ | /g/ | b | 471 | 30.8 | -146 | 6.2 | 617 |
| F2i=2000 | | | | d | 220 | 12.4 | -325 | 12.9 | 545 |
| F3i=2200 | | | | g | 182 | 10.0 | -291 | 11.7 | 473 |
| Ambiguous [b-d] | /b/ | /g/ | /b/ | b | -29 | 1.9 | 154 | 6.6 | 183 |
| F2i=1500 | | | | d | -280 | 15.7 | 25 | 1.0 | 305 |
| F3i=2500 | | | | g | -318 | 17.5 | 9 | 0.4 | 327 |
| Ambiguous [d-g] | /g/ | /b/ | /g/ | b | 371 | 24.3 | 54 | 2.3 | 425 |
| F2i=1900 | | | | d | 120 | 6.7 | -125 | 4.95 | 245 |
| F3i=2400 | | | | g | 82 | 4.5 | -91 | 3.7 | 173 |
| Ambiguous [b-d-g] | /d/ | /b/ | /b/ | b | 171 | 11.2 | -146 | 6.2 | 317 |
| F2i=1700 | | | | d | -80 | 4.5 | -325 | 12.9 | 405 |
| F3i=2200 | | | | g | -118 | 6.5 | -291 | 11.7 | 409 |

close to 930 Hz and F3i close to 2346 Hz would provide a clear /b/; in fact, we find no majority /b/ responses in this whole set. The stimulus with the closest formant requirements for predicted *b* in fact yields either a majority /g/ response or an ambiguous [d-g] reponse. Clearly, subjects are not solely relying on formant patterns from measurement, as reflected in the regression model.

However, the stimulus onset values that were used here are not actually found in actual measurements of stops (see Chapter 2 where none of the measured data had such low onset values.) Thus, the /b/ predictions for stimuli having these low onset values were based on an extrapolation of the /b/ regression line. It is also interesting to note that where there was a 'gap' of F3i vs. F3v measured values for /b/ stops (i.e., at low F3i and F3v levels), subjects respond with majority /g/ responses; thus, the scarcity of /b/ responses may be related to the lack of low F3 values for /b/'s in natural data. [5] (The scarcity of /b/ responses in general will be more fully discussed below.)

The /g/ responses for this set also do not coincide with predicted formant values of *g*. Three examples of 'good' /g/'s are shown in Figure 3.4. Two of these /g/'s have F2i and F3i values that lie closest to the *b* regression line (i.e., category response would be predicted to be b). The other 'good' /g/ has F2i value closest to *d*, and F3i value closest to *b*. However, the second choice predicted by both F2i vs. F2v and F3i vs. F3v for all three examples of 'good' /g/ is *g*.

---

[5] More complex models of the distribution of natural data might hold some promise. Such models, however, are beyond the scope of the present work.

The F2i of 'ambiguous' [d-g] is closest to predicted *g*, though this value does fall between the predicted *d* and *g* values.

F3i predictions for all of these stimuli favor /b/. Thus for this vowel set (F2v=900), correct predictions based on F2i formant measurments are obtained only for 'good' /d/ and 'ambiguous' [d-g], while predictions based on F3i formant measurements are obtained only for the /b/ stimuli.

Predictions based on both F2 and F3 information combined are also presented in Table 3.2 for this vowel set. The absolute values of the deviations in Hz from both F2i vs. F2v and F3i vs. F3v are summed; the category for which the smallest value is obtained is considered to be the 'predicted' category. Results show that, in all cases for this vowel set, the predicted category from F2i information alone and the predicted category from F2i and F3i information combined are identical.

### F2v=1050 Hz (Ambiguous Vowel [o-U])

For F2v=1050, we would predict that a stimulus with F2i=1062 Hz and F3i=2346 Hz would obtain a majority /b/ response; again we find few /b/ responses for any of these stimuli. The only stimulus that showed some /b/ responses was the stimulus having an F2i value of 1300 Hz and F3i value of 2100 Hz and was ambiguously identified as either /b/, /d/ or

/g/; for this case, F2i is closest to predicted F2i of *g*, and F3i is closest to (although lower than) the predicted F3i of *b*.

The ambiguous [d-g] stimulus had F2i=1400 Hz, which is closest to predicted *d*, although it lies between the predicted *d* and *g* values. again, the ambiguity may be due to values which are between two regression lines.

The 'good /g/' stimulus had F2i=1000 Hz, which is closest to predicted *b*, and F3i=1700 Hz, which is much lower than any of the predicted F3i values for *b*, *d*, or *g* (but is closest to *b*). For this vowel set, as in the previous set, *b* predictions are heard as /g/'s, though second choice predictions are for /g/'s. The 'good /d/' in this series had F2i=1400 Hz and F3i=2400 Hz; here, F2i is closest to the predicted F2i value of *d*.

Again, F3i predictions favor /b/ responses for all these stimuli. However, second choice predictions from F3i information (as indicated by the second smallest value in deviations from the regression lines) are for /g/'s.

Thus for this set of stimuli (F2v=1050 Hz), predicted F2i values are obtained for the good /d/ and ambiguous [d-g] cases, but are not obtained for the /b/ or /g/ responses. As in the the previous F2v set, F3i predictions are obtained only for the /b/ stimuli.

Predictions for F2i and F3i combined (shown in Table 3.3) are identical to F2i predictions for all stimuli in this vowel set.

F2v=1200 Hz (Vowel /U/)

For F2v=1200 Hz, /b/ is predicted for a stimulus with F2i near 1190 Hz and F3i near 2347 Hz; the stimulus value with F2i=1100 Hz and F3i=2300 Hz in fact gets over 70% /b/ responses, which closely follows the expected result.

The 'good' /d/ stimulus has F2i=1700 Hz which is closest to the predicted F2i value of *d*. The good /g/ stimulus (F2i=1300 Hz and F3i=1700 Hz) has both F2i and F3i closest to the predicted *b* values. Thus, for the 'good' stimuli, F2i predictions are correct for /b/ and /d/, but not for /g/ and F3i predictions are only correct for /b/.

The F2i of the 'ambiguous' [d-g] stimulus is closest to predicted *d*. The 'ambiguous' [b-g] stimulus (F2i=900 Hz, F3i=2400 Hz) has F2i and F3i values closest to predicted *b*. The 'ambiguous' [b-d-g] stimulus (F2i=1300 Hz, F3i=2000 Hz) also has F2i and F3i values closest to predicted *b*. Here, the 'ambiguous' stimuli show at least one of the response categories predicted by F2i while F3i predictions again consistently predict a /b/ response for all of the stimuli. Thus the F2i vs. F2v regression lines (of Figure 2.8) seem to make better predictions of category response than do the F3i vs. F3v regression lines.

Predictions from F2i and F3i combined are identical to the predictions of F2i. It is interesting to note that for all vowel sets with lower F2v's (i.e., F2v=1200 Hz or less), the combined F2i-F3i predictions were identical to F2i predictions alone.

### F2v=1450 Hz (Ambiguous [U-ε] Vowel)

For F2v=1450 Hz, the F2i of 'good' /b/ (F2i=1200 Hz), lies closest to predicted *b*. However, the 'good' /d/ is predicted to be *g* and the 'good' /g/ is predicted to be *d*. However, the *d* and *g* predictions are very close to each other here (only 46 Hz apart as shown in Table 3.1). As in the previous vowel sets, F3i predictions consistently favor /b/'s.

The F2i of 'ambiguous' [b-d] (F2i=1500 Hz) lies closest to predicted F2i of *b* but lies between *b* and *d*. Similarly, the F2i of 'ambiguous' [b-d-g] (F2i=1500 Hz) lies closest to predicted *b* but lies between *b* and *g*). Thus, for this set of F2v values, reasonable F2i predictions are supported for the 'good' /b/ stimulus and the 'ambiguous' [b-d] stimulus, but the F3i predictions are not good.

Predictions from F2i and F3i combined are shown in Table 3.5. The only predictions which are different from predictions obtained with F2i alone are for the 'good' /g/ and ambiguous [d-g] stimuli. Combined F2i and F3i predicts the 'good' /g/ stimulus would be ambiguous between /d/ and /g/, whereas the F2i predicts that subjects would hear this stimulus as a /d/. Thus the combined prediction seems to be a slightly more appropriate measure for this particular case. However, the 'ambiguous [d-g] stimulus is predicted by F2i to be heard as a /g/, but the combined F2i and F3i prediction is that subjects would hear this stimulus as a /b/. Thus, for this case, the F2i prediction alone seems to be slightly more appropriate.

### F2v=1650 Hz (Vowel /ɛ/)

For F2v=1650 Hz, the F3i vs. F3v lines make better predictions than for the other F2v sets. In fact, this is the only set of stimuli where both F2i and F3i make fairly good predictions. The 'good' /b/ stimulus (F2i=1400 Hz, F3i=2200 Hz) has both F2i and F3i closest to predicted *b*. The 'good' /d/ stimulus (F2i=1700 Hz, F3i=2600 Hz) has both F2i and F3i closest to predicted d. The 'good' /g/ stimulus (F2i=2000 Hz, F3i=2200 Hz) has F2i closest to predicted *g*, but F3i closest to predicted *b*.

The 'ambiguous' [b-d] stimulus (F2i=1500 Hz, F3i=2500 Hz) has F2i closest to predicted *b*, and F3i is closest to predicted *g* (but falls between *d* and *g*). The 'ambiguous' [d-g] stimulus (F2i=1900 Hz, F3i=2400 Hz) has F2i closest to predicted *g*. The 'ambiguous' [b-d-g] stimulus (F2i=1700 Hz, F3i=2200 Hz) has F2i closest to predicted *d* (though it falls between *d* and *b*, and F3i closest to predicted *b*.

Predictions from combined F2i and F3i are identical to predictions from F2i alone except for the 'ambiguous' [b-d-g] stimulus; here, F2i predicts that subjects will hear a /d/, while the combined F2i and F3i prediction is that subjects will hear a /b/. For this case, it is difficult to say which of the two types of predictions are more appropriate, since the actual stimulus was ambiguously heard as either /b/, /d/ or /g/.

### 3.2.5.1 Summary of Formant Information

Formant information, at least as represented by the regression lines of Figure 2.8, is clearly not the only information that listeners attend to when they must categorize the place of articulation in stop consonants. Considered separately, F2i vs. F2v predictions are clearly better than F3i vs. F3v predictions. For low F2v's, the F3i predictions consistently predicted a /b/ category response, whereas listeners responded with other categories (mainly /g/'s). The formant predictions for F2i and F3i are very good only at high F2 steady-state values (e.g., F2v=1650 Hz.)

When F2i and F3i information is combined, the predictions are almost identical to the predictions obtained from F2i alone. It is thus reasonable to suppose that, if subjects are relying on formant relationships as represented by the regression lines, F2i would be a 'stronger' cue than F3i (at least for these stimuli). In fact, F2i made many more correct predictions than did F3i, when F2i and F3i were considered separately.

This study showed the extent to which a formant-based prediction of consonant category responses is correct. The specific failures of formant-based information were also addressed. In the following section, the role of spectral shape information is discussed.

### 3.2.6 The Use of Spectral Shape Information

The role of spectral information will be discussed in terms of three previous studies:

1. the Steven's and Blumstein templates (1979)
2. Lahiri and Blumstein templates (1981) and
3. Kewley-Port features (1980).

#### 3.2.6.1 Stevens and Blumstein Templates

Tables 3.7-3.11 show the predicted vs. obtained category responses using Steven's and Blumstein's onset spectral shape templates for vowels /o/, [o-U], /U/, [U-ɛ], and /ɛ/, respectively. A /b/ response is predicted when the onset spectral shape is diffuse falling; a /d/ response is predicted when the onset spectral shape is diffuse rising; a /g/ response is predicted when the onset spectral shape is compact in the mid-frequency range.

The Stevens and Blumstein templates were fit to the onset spectra in Figures 3.4-3.8. If more than one template fit an onset spectrum, that stimulus was considered to have an 'ambiguous' prediction.

The results in Tables 3.7-3.11 show that these onset spectral-shape templates made good predictions only for the stimuli set which had high F2v's. For stimuli having lower F2v's, these templates correctly predicted 'good' /g/'s when the formant onsets of these stimuli were close together (less than 500 Hz apart). However, it is important to keep in mind that formant

5

## TABLE 3.7

## Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment; Vowel /o/ (F2v=900 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /g/<br>F2i=700<br>F3i=1600 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /g/<br>F2i=1000<br>F3i=1500 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /g/<br>F2i=1600<br>F3i=1700 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Good /d/<br>F2i=1400<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Ambiguous [d-g]<br>F2i=1100<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |

## TABLE 3.8

## Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment; Vowel [o-U] (F2v=1050 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /d/ | /b/ | - compact |
| F2i=1400 | | - diffuse rising |
| F3i=2400 | | + diffuse falling |
| Good /g/ | /b/ | - compact |
| F2i=1000 | | - diffuse rising |
| F3i=1700 | | + diffuse falling |
| Ambiguous [d-g] | /g/ | + compact |
| F2i=1400 | | - diffuse rising |
| F3i=1700 | | - diffuse falling |
| Ambiguous [b-d-g] | /b/ | - compact |
| F2i=1300 | | - diffuse rising |
| F3i=2100 | | + diffuse falling |

## TABLE 3.9

## Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment; Vowel /U/ (F2v=1200 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>(70%)<br>F2i=1100<br>F3i=2300 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /d/<br>F2i=1700<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /g/<br>F2i=1300<br>F3i=1700 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Ambiguous<br>[d-g]<br>F2i=1700<br>F3i=2000 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Ambiguous<br>[b-g]<br>F2i=900<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Ambiguous<br>[b-d-g]<br>F2i=1300<br>F3i=2000 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |

## TABLE 3.10

## Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment; Vowel [U-ɛ] (F2v=1450 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>F2i=1200<br>F3i=2000 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /d/<br>F2i=1600<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /g/<br>F2i=1700<br>F3i=1900 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Ambiguous<br>[d-b]<br>F2i=1500<br>F3i=2300 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Ambiguous<br>[d-g]<br>F2i=1600<br>F3i=2000 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Ambiguous<br>[b-d-g]<br>F2i=1500<br>F3i=1900 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |

## TABLE 3.11

## Obtained vs. Predicted Response Categories From Stevens and Blumstein Onset Spectral-Shape Templates; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>F2i=1400<br>F3i=2200 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Good /d/<br>F2i=1700<br>F3i=2600 | /d/ | - compact<br>+ diffuse rising<br>- diffuse falling |
| Good /g/<br>F2i=2000<br>F3i=2200 | /g/ | + compact<br>- diffuse rising<br>- diffuse falling |
| Ambiguous<br>[b-d]<br>F2i=1500<br>F3i=2500 | [b-d] | - compact<br>+ diffuse rising<br>+ diffuse falling |
| Ambiguous<br>[d-g]<br>F2i=1900<br>F3i=2400 | /b/ | - compact<br>- diffuse rising<br>+ diffuse falling |
| Ambiguous<br>[b-d-g]<br>F2i=1700<br>F3i=2200 | [b-d-g] | - compact<br>- diffuse rising<br>- diffuse falling |

onsets and onset spectral shape are not independent in cascade synthesis. The F2 and F3 formant transitions were manipulated independently, but the onset spectral shape was not an independent variable. The diffuse falling templates fit many of the stimuli, since the stimuli were synthesized with voicing at stimulus onset; owing to the cascade synthesis technique and the glottal source spectrum, this generally resulted in falling spectra at stimulus onset. It should be noted that only two synthetic stimuli fit the diffuse rising template (for F2v=1600 Hz). If 'diffuse rising' (as defined by Stevens and Blumstein templates) is a cue for /d/'s, none of these synthetic stimuli at lower F2v's should have been heard as /d/'s, since they had diffuse falling spectra. However, majority responses to /d/ were found even when the onset spectral shape was diffuse falling.

It is interesting to note that stimuli which were categorized as 'good' /d/'s were often predicted to be /d/'s by formant parameters, but were predicted to be /b/'s by the templates. Thus, formant parameters seem to make better predictions for 'good' /d/'s than the Stevens and Blumstein templates.

### 3.2.6.2 Lahiri and Blumstein Ratios

Tables 3.12-3.16 show the obtained and predicted response categories from the ratio values proposed by Lahiri and Blumstein for vowels /o/, [o-U], /U/, [U-ε], and /ε/, respectively. Lahiri and Blumstein used a

## TABLE 3.12

### Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /o/ (F2v=900 Hz )

| Obtained Category | Predicted Category | Ratio Value |
|---|---|---|
| Good /g/<br>F2i=700<br>F3i=1600 | /b/ | 1.16 |
| Good /g/<br>F2i=1000<br>F3i=1500 | /b/ | .89 |
| Good /g/<br>F2i=1600<br>F3i=1700 | /b/ | 4.0 |
| Good /d/<br>F2i=1400<br>F3i=2400 | /b/ | .79 |
| Ambiguous [d-g]<br>F2i=1100<br>F3i=2400 | /b/ | 1.75 |

## TABLE 3.13

## Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel [o-U] (F2v=1050 Hz )

| Obtained Category | Predicted Category | Ratio Value |
|---|---|---|
| Good /d/<br>F2i=1400<br>F3i=2400 | /b/ | .93 |
| Good /g/<br>F2i=1000<br>F3i=1700 | /b/ | 1.27 |
| Ambiguous [d-g]<br>F2i=1400<br>F3i=1700 | /b/ | 1.42 |
| Ambiguous [b-d-g]<br>F2i=1300<br>F3i=2100 | /b/ | .76 |

## TABLE 3.14

## Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /U/ (F2v=1200 Hz )

| Obtained Category | Predicted Category | Ratio Value |
|---|---|---|
| Good /b/<br>F2i=1100<br>F3i=2300 | /b/ | 1.21 |
| Good /d/<br>F2i=1700<br>F3i=2400 | /b/ | .63 |
| Good /g/<br>F2i=1300<br>F3i=1700 | /b/ | 1.27 |
| Ambiguous [d-g]<br>F2i=1700<br>F3i=2000 | /b/ | 1.11 |
| Ambiguous [b-g]<br>F2i=900<br>F3i=2400 | /b/ | .82 |
| Ambiguous [b-d-g]<br>F2i=1300<br>F3i=2000 | /b/ | 1.13 |

## TABLE 3.15

### Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel [U-ɛ] (F2v=1450 Hz )

| Obtained Category | Predicted Category | Ratio Value |
|---|---|---|
| Good /b/<br>F2i=1200<br>F3i=2000 | /b/ | 1.24 |
| Good /d/<br>F2i=1600<br>F3i=2400 | /b/ | .65 |
| Good /g/<br>F2i=1700<br>F3i=1900 | /b/ | 2.71 |
| Ambiguous [d-b]<br>F2i=1500<br>F3i=2300 | /b/ | .88 |
| Ambiguous [d-g]<br>F2i=1600<br>F3i=2000 | /b/ | 1.0 |
| Ambiguous [b-d-g]<br>F2i=1500<br>F3i=1900 | /b/ | 1.5 |

## TABLE 3.16

## Obtained vs. Predicted Response Categories From Lahiri and Blumstein Features; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz )

| Obtained Category | Predicted Category | Ratio Value |
|---|---|---|
| Good /b/<br>F2i=1400<br>F3i=2200 | /b/ | .88 |
| Good /d/<br>F2i=1700<br>F3i=2600 | /d/ | .48 |
| Good /g/<br>F2i=2000<br>F3i=2200 | /b/ | 1.44 |
| Ambiguous [b-d]<br>F2i=1500<br>F3i=2500 | /b/ | .65 |
| Ambiguous [d-g]<br>F2i=1900<br>F3i=2400 | /b/ | .80 |
| Ambiguous [b-d-g]<br>F2i=1700<br>F3i=2200 | /b/ | .71 |

revised definition of 'diffuse falling' and 'diffuse rising' (though they did not discuss the feature 'compact'). This revision was done in order to account for French alveloar and Malayalam dental and alveolar stops. Alveolars, they suggested, had a 'predominance of high frequency energy at the onset relative to the later-occurring energy distribution at the onset of voicing', while for labials, 'there does not seem to be a change in the relative distribution of energy in the high and low frequencies from the burst release to the onset of voicing'. They 'drew a line' between the F2 and F4 peaks of the burst spectrum and the spectrum at voicing onset; the ratio of the difference in energy from the burst to the onset of voicing was then calculated, using 1500 Hz for the 'low' frequency and 3500 Hz for the 'high' frequency. If the ratio was greater than .5, this indicated a labial consonant. An alveolar consonant had a ratio less than .5 (or a negative value).

Although the synthetic stimuli did not have a distinct burst and voicing onset was always at stimulus onset, the Lahiri and Blumstein ratios were calculated for these stimuli using the onset spectrum as the 'burst' spectrum and the specrum after four frames (20 msec) as the 'voicing onset' spectrum. It was felt that if these ratios are important cues, they should be apparent in synthetic stimuli as well as natural stimuli; subjects, after all, were able to categorize the synthetic stimuli, so it is not unreasonable to check ratio values for these stimuli. These ratios are really only designated as alveolar and labial cues; however, all of the well and poorly identified stimuli were analyzed, including velars.

As can be seen in Tables 3.12 - 3.16, these ratios correctly predict only one 'good' /d/ stimulus, and make /b/ predictions for all other stimuli. Again, the high ratios reflect the falling spectral patterns of these synthetic stimuli, due to the voicing excitation and low F1 at stimulus onset, resulting in a falling spectral shape due to the cascade synthesis technique. It was felt that if these ratios were important cues to consonant categorization, the proper ratios for 'good' tokens should have been evident.

3.2.6.3 Kewley-Port Features

Tables 3.17-3.21 show predicted vs. obtained response categories from Kewley-Port's features (Kewley-Port, 1980), for vowels /o/, [o-U], /U/, [U-ɛ], and /ɛ/, respectively. These features are:

1. rising vs. falling onset spectral shape,
2. late onset of low frequency energy and
3. mid-frequency peaks over time.

'Late onset of low frequency energy' is essentially a measure of VOT (voice onset time). Since all the synthetic stimuli had 'voicing' immediately at stimulus

## TABLE 3.17

## Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel /o/ (F2v=900 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /g/<br>F2i=700<br>F3i=1600 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=1000<br>F3i=1500 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=1600<br>F3i=1700 | /g/ | + Falling spectrum<br>+ Late onset of low frequency energy<br>+ Mid-frequency peaks over time |
| Good /d/<br>F2i=1400<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[d-g]<br>F2i=1100<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |

### TABLE 3.18
### Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel [o-U] (F2v=1050 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /d/<br>F2i=1400<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=1000<br>F3i=1700 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[d-g]<br>F2i=1400<br>F3i=1700 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[b-d-g]<br>F2i=1300<br>F3i=2100 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |

TABLE 3.19

Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment;

Vowel /U/ (F2v=1200 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>(70%)<br>F2i=1100<br>F3i=2300 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /d/<br>F2i=1700<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=1300<br>F3i=1700 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[d-g]<br>F2i=1700<br>F3i=2000 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[b-g]<br>F2i=900<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[b-d-g]<br>F2i=1300<br>F3i=2000 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |

## TABLE 3.20
## Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel [U-ε] (F2v=1450 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>F2i=1200<br>F3i=2000 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /d/<br>F2i=1600<br>F3i=2400 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=1700<br>F3i=1900 | /g/ | + Falling spectrum<br>- Late onset of low frequency energy<br>+ Mid-frequency peaks over time |
| Ambiguous<br>[d-b]<br>F2i=1500<br>F3i=2300 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[d-g]<br>F2i=1600<br>F3i=2000 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[b-d-g]<br>F2i=1500<br>F3i=1900 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |

### TABLE 3.21

### Obtained vs. Predicted Response Categories From Kewley-Port Features; F2v Experiment; Vowel /ɛ/ (F2v=1650 Hz )

| Obtained Category | Predicted Category | Features |
|---|---|---|
| Good /b/<br>F2i=1400<br>F3i=2200 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /d/<br>F2i=1700<br>F3i=2600 | /d/ | + Rising spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Good /g/<br>F2i=2000<br>F3i=2200 | /g/ | + Falling spectrum<br>- Late onset of low frequency energy<br>+ Mid-frequency peaks over time |
| Ambiguous<br>[b-d]<br>F2i=1500<br>F3i=2500 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[d-g]<br>F2i=1900<br>F3i=2400 | [b-d] | (+ Flat spectrum)<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |
| Ambiguous<br>[b-d-g]<br>F2i=1700<br>F3i=2200 | /b/ | + Falling spectrum<br>- Late onset of low frequency energy<br>- Mid-frequency peaks over time |

onset, this feature did not vary.

Results in Tables 3.17-3.21 show that the Kewley-Port spectral features make good predictions only for higher F2v's (Table 3.21). For lower F2v's, these features consistently predict /b/ responses for the synthetic stimuli having a 'diffuse' onset spectrum. The amplitude synthesis parameters forced a falling onset spectrum; nevertheless, subjects identified some of these stimuli as 'good' /d/'s, even though the spectrum was not 'rising'. It was felt that if this were a /d/ cue, subjects would not have responded with majority /d/ responses for those stimuli having a diffuse falling spectrum. The Kewley-Port features also predicted /g/'s when F2i and F3i were very close together (less than 200 Hz apart). However, subjects reponded with majority /g/ responses for stimuli whose F2i and F3i were much farther apart than 200 Hz (e.g., for stimuli which had F2v=900 Hz, one 'good' /g/ stimulus had F2i=700 Hz and F3i=1600 Hz, and another 'good' /g/ stimulus had F2i=1000 Hz and F3i=1500 Hz). In addition, some of the stimuli which were categorized as 'good' /g/'s had a prominent mid-frequency peak only at stimulus onset, and this peak did not 'extend through time.' Again, if this feature was a necessary cue for /g/, subjects should not have responded with a majority /g/ response if the peak did not extend through time.

### 3.2.7 Summary and Comparison

Tables 3.22-3.26 show a summary and comparison of the formant and spectral predictions for the vowels /o/, [o-U], /U/, [U-ε], and /ε/, respectively. Results show that both formant and spectral shape information are needed and, furthermore, a redefinition of some of the spectral shape cues seems to be necessary. It is also of interest to see where formant and spectral predictions make similar predictions, and where they make differing predictions. In those cases where different predictions are made by formant and spectral shape parameters, it is of interest to see which parameter set is more appropriate (i.e., which parameters correctly predict the categories which were obtained in the perception experiment.

A 'correct prediction' of a Stevens and Blumstein template indicates that a template appropriate to the majority response fit the onset spectral shape of the stimulus while no other template falsely accepted that stimulus. The Lahiri and Blumstein ratios are considered to make correct predictions only if the ratio values are above .5 for stimuli with a majority response of /d/, and below .5 for stimuli heard as /b/. Kewley-Port features are considered to make correct predictions if the spectral shape features are appropriate (as defined by Kewley-Port) to subjects' dominant responses. Finally, 'correct predictions' of the formant-based regression lines (of Figure 2.8) indicates that the stimulus' values were closer to the

## TABLE 3.22

## Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /o/

| Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|
| | F2i vs. F2v | F3i vs. F3v | Stevens & Blumstein | Lahiri & Blumstein | Kewley-Port |
| Good /g/<br>F2i=700<br>F3i=1600 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Good /g/<br>F2i=1000<br>F3i=1500 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Good /g/<br>F2i=1600<br>F3i=1700 | /d/ | /b/ | /g/ | /b/ | /g/ |
| Good /d/<br>F2i=1400<br>F3i=2400 | /d/ | /b/ | /b/ | /b/ | /b/ |
| Ambiguous<br>[d-g]<br>F2i=1100<br>F3i=2400 | /g/ | /b/ | /b/ | /b/ | /b/ |

## TABLE 3.23

## Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel [o-U]

| Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|
| | F2i vs. F2v | F3i vs. F3v | Stevens & Blumstein | Lahiri & Blumstein | Kewley-Port |
| Good /d/<br>F2i=1400<br>F3i=2400 | /d/ | /b/ | /b/ | /b/ | /b/ |
| Good /g/<br>F2i=1000<br>F3i=1700 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Ambiguous<br>[d-g]<br>F2i=1400<br>F3i=1700 | /d/ | /b/ | /g/ | /b/ | /b/ |
| Ambiguous<br>[b-d-g]<br>F2i=1300<br>F3i=2100 | /g/ | /b/ | /b/ | /b/ | /b/ |

## TABLE 3.24

## Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /U/

| Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|
| | F2i vs. F2v | F3i vs. F3v | Stevens & Blumstein | Lahiri & Blumstein | Kewley-Port |
| Good /b/ (70%) F2i=1100 F3i=2300 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Good /d/ F2i=1700 F3i=2400 | /d/ | /b/ | /b/ | /b/ | /b/ |
| Good /g/ F2i=1300 F3i=1700 | /b/ | /b/ | /g/ | /b/ | /b/ |
| Ambiguous [d-g] F2i=1700 F3i=2000 | /d/ | /b/ | /g/ | /b/ | /b/ |
| Ambiguous [b-g] F2i=900 F3i=2400 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Ambiguous [b-d-g] F2i=1300 F3i=2000 | /b/ | /b/ | /b/ | /b/ | /b/ |

### TABLE 3.25

### Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel [U-ɛ]

| Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|
| | F2i vs. F2v | F3i vs. F3v | Stevens & Blumstein | Lahiri & Blumstein | Kewley-Port |
| Good /b/<br>F2i=1200<br>F3i=2000 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Good /d/<br>F2i=1600<br>F3i=2400 | /g/ | /b/ | /b/ | /b/ | /b/ |
| Good /g/<br>F2i=1700<br>F3i=1900 | /d/ | /b/ | /g/ | /b/ | /g/ |
| Ambiguous<br>[b-d]<br>F2i=1500<br>F3i=2300 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Ambiguous<br>[d-g]<br>F2i=1600<br>F3i=2000 | /g/ | /b/ | /g/ | /b/ | /b/ |
| Ambiguous<br>[b-d-g]<br>F2i=1500<br>F3i=1900 | /b/ | /b/ | /g/ | /b/ | /b/ |

## TABLE 3.26

## Summary and Comparisons of Obtained and Predicted Response Categories; F2v Experiment; Vowel /ɛ/

| Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|
| | F2i vs. F2v | F3i vs. F3v | Stevens & Blumstein | Lahiri & Blumstein | Kewley-Port |
| Good /b/<br>F2i=1400<br>F3i=2200 | /b/ | /b/ | /b/ | /b/ | /b/ |
| Good /d/<br>F2i=1700<br>F3i=2600 | /d/ | /d/ | /d/ | /d/ | /d/ |
| Good /g/<br>F2i=2000<br>F3i=2200 | /g/ | /g/ | /g/ | /b/ | /g/ |
| Ambiguous<br>[b-d]<br>F2i=1500<br>F3i=2500 | /b/ | /g/ | [b-d] | /b/ | /b/ |
| Ambiguous<br>[d-g]<br>F2i=1900<br>F3i=2400 | /g/ | /b/ | /b/ | /b/ | [b-d] |
| Ambiguous<br>[b-d-g]<br>F2i=1700<br>F3i=2200 | /d/ | /b/ | [b-d-g] | /b/ | /b/ |

appropriate regression line than to any of the other regression lines.

Vowel /o/

For the vowel /o/ (Table 3.22), the Stevens and Blumstein compact template correctly identified only one of the 'good' /g/ stimuli; the other 'good' /g/'s were incorrectly predicted to be /b/'s by the spectral shape feature sets (i.e., fit only the diffuse falling template). The formant onset values of these stimuli were closer to either the /b/ and /d/ regression lines than to the /g/ regression line. The 'good' /d/ stimulus, on the other hand, was closest to the /d/ regression lines (for F2i vs. F2v or when F2i and F3i were combined). The 'ambiguous' [d-g] stimulus was closest to the /g/ F2i vs. F2v regression line, but was accepted by the /b/ spectral shape parameters.

It should be noted than none of the parameter sets predict both well and poorly identified stimuli for all consonant categories, and that neither the formant-based regression lines nor spectral shape parameters correctly predicted subjects' majority /g/ responses. Specifically, the Stevens and Blumstein compact templates correctly predicted one of the 'good' /g/ stimuli (i.e., was correctly accepted by the compact template), while F2i formant information correctly predicted the 'good' /d/ stimulus (i.e., the 'good' /d/ stimulus was closest to the F2i vs. F2v /d/ regression line).

As was noted above, this stimulus set did not receive majority /b/ responses, which was an unexpected result. Why did subjects consistently label predicted /b/ stimuli (predicted by both formant and spectral shape parameters) as /g/'s? Some plausible explanations for the lack of /b/ reponses are as follows:

1. Perhaps the overall amplitude characteristics of of these stimuli were too high at stimulus onset, which would cue alveolar or velar consonants (cf. Repp, 1978). Alternatively, perhaps other aspects of the synthetic stimuli of Stevens and Blumstein (which were manipulated differently depending on the consonant) cued more /b/ responses; such aspects include rate of formant transitions and F1 values.
2. Perhaps subjects were not responding with many /b/'s in this vowel set because the F3i values were lower than is normally found in natural data. Recall that in natural measurement, there was a 'gap' of low F3i values for measured /b/'s, except when F3v was low (i.e., before the vowel /ɚ/). In measurements, /g/'s before vowels near this F2v value (F2v=900 Hz) had low F3i's. Thus, perhaps subjects were responding to the overall distribution patterns of formant measurements found in natural data (cf. Nearey, Hogan and Roszypal, 1979).
3. The 'compact' onset spectral shape was not compact at mid-frequencies, but rather was compact at low frequencies. Thus a velar categorization was obtained,

since there was compactness towards the F1 range (instead of compactness in the F2 and F3 range). This extension of the range of compactness would be an interesting feature to test in further investigations.

4. An additional aspect of the compact feature that would be interesting to test is an extension of the bandwidths of the proposed compact templates. For the 'good' /g/ stimulus with F2i=1000 Hz and F3i=1500 Hz (Figure 3.4), none of the compact templates fit the onset spectrum. However, the onset spectrum has a 'compactness' in the sense that there is significant energy in a limited band of frequencies (within 700-1500 Hz) and above this band, there are no significant energy peaks.

### Vowel [o-U]

For the vowel [o-U] (Table 3.23), the 'good' /d/ was closest to the /d/ regression lines. 'Good' /g/'s are again predicted to be /b/'s by both spectral shape and formant parameters (i.e., are accepted by the diffuse falling template, and lie closest to the /b/ regression lines). The 'ambiguous' [d-g] stimulus was predicted to be /g/ by the Stevens and Blumstein templates, and /d/ from the F2 regression lines. For this case then, the ambiguity of the signal may be a result of a conflict in formant and spectral shape cues. As in the vowel /o/ set, both spectral and formant information seem to be playing a perceptual role.

Again, predicted /b/'s are categorized as /g/'s by subjects; a redefinition of 'compact' as outlined above also seems in order for this vowel set.

Vowel/U/

For the vowel /U/, (Table 3.24), both spectral shape and formant parameters correctly predict the 'good' /b/ stimulus. The 'good' /d/ stimulus lies closer to the /d/ regression lines than to the other regression lines. The 'good' /g/ stimulus fits the appropriate compact Stevens and Blumstein template. The ambiguous [d-g] stimulus is predicted to be /d/ by the regression lines and /g/ by the Stevens and Blumstein spectral templates. Perhaps the ambiguity is a result of 'conflicting' spectral shape and formant cues. The other ambiguous stimuli are predicted to be /b/ by all models.

Vowel [U-ɛ]

For the vowel [U-ɛ] (Table 3.25), the 'good' /b/ is correctly predicted by both spectral and formant information. The 'good' /d/ was closest to the /g/ F2 regression line (cf. Table 3.5), but the /d/ and /g/ regression lines of F2i vs. F2v were very close, owing to the '/d/-/g/ crossover' near this F2v value (see Figure 2.8). The 'good' /g/ stimulus had the appropriate compact spectral shape at stimulus onset.

Vowel /ɛ/

For the vowel /ɛ/ (Table 3.26), both formant and spectral parameters seemed to make fairly good predictions of the subjects' response categories. In particular, F2 information, the Stevens and Blumstein (spectral) onset templates and the Kewley-Port (spectral) features correctly predict all 'good' tokens. Thus, both formant and spectral models make the same predictions.

### 3.2.7.1 Conclusions

None of the models made correct consonant category predictions for every vowel set. The 'good' /b/ stimuli were predicted to be /b/ by both formant and spectral-shape parameters. However, /b/'s were only reliably heard at higher F2v's; for these higher F2v's, both the formant and Steven's and Blumstein templates correctly predicted these stimuli to be /b/. However, at lower F2v's, stimuli predicted to be /b/ by both formant and spectral-shape parameter sets were not heard as /b/, but rather, were heard as /g/. Formant predictions (based on the regression lines of Figure 2.8) made more correct /d/ predictions than did spectral-shape predictions; of the five 'good' /d/ stimuli four were closer to the /d/ regression lines than to the other regression lines, but only one of the five 'good' /d/ stimuli had the appropriate diffuse rising onset spectral shape. Finally, four out of the seven 'good' /g/ stimuli fit the appropriate compact template, but only one of these stimuli was closest to the /g/ regression lines. These results seem to suggest that formant-based information is useful for predicting /d/'s and spectral-shape information is useful for predicting /g/'s.

What was the role of F2v? For low F2v's, 'compactness' of the onset spectrum seemed to cue /g/'s, while formant information seemed to cue /d/'s; /b/'s were not reliably heard at low F2v's, contrary to both spectral shape and formant-based predictions. Thus, at low F2v's, formant and spectral information made *different* predictions. At intermediate F2v's, however, *both* formant and spectral-shape information made correct predictions for the /b/ category. However, for these F2v's, spectral shape and formant information made different predictions for /d/'s and /g/'s; again, formant predictions seemed to be better for /d/'s and spectral shape information seemed to make better predictions for /g/'s. At the highest F2v, however, both formant and spectral-shape information made the *same* (correct) predictions for *all* of the 'good' consonant categories.

In this chapter, *acoustic-phonetic* features were analyzed. The roles of formant information and spectral onset information were discussed, and it was felt that *both* parameters were important cues for listeners. In the following chapter, specific interactions of both acoustic and phonetic factors will be addressed.

# 4. PHONETIC AND ACOUSTIC INTERACTIONS IN CONSONANT CATEGORIZATIONS

In the last chapter, it was shown that acoustic context was important. Since the acoustic context influenced subjects' consonant categorizations, it can be argued that the 'unit of perception' must be context sensitive. What is the nature of this context sensitivity? Does the perceptual system require a phonetic label of the vowel context, or does it only require some acoustic information (e.g., F2v or F3v)? Is this information required before a consonant categorization decision is made (i.e, 'hierarchical'), or are the consonant and vowel decisions made simultaneously (i.e., via a syllabic template)?

There are four types of perceptual units which can be described:

1. context-free, acoustic based segment units
2. context-sensitive, acoustic-based segment units (i.e., acoustically tuned),
3. context sensitive, phonetically based syllable units, whereby decisions are made hierarchically and
4. context sensitive, phonetically based syllable units, whereby decisions are made simultaneously (i.e., syllable templates).

In this chapter the role of context sensitivity is discussed. Are subjects' context-sensitive responses based on prior 'vowel label' decisions, or can their responses be described in terms of acoustic factors in the following vowel? The aspect of 'hierarchical' vs. 'simultaneous' decisions will not be addressed. What is of interest here is the 'level of processing': are subjects responding on the basis of the acoustics of the vowel context or are they responding on the basis of the vowel label? In the main perception experiment of Chapter 3, subjects were also asked to identify the vowel of the CV synthetic stimuli. Since the F2v continuum also contained 'ambiguous' [o-U] and [U-ɛ] vowels, the role of vowel labelling could be investigated. A continuum of formant onsets for these 'ambiguous' vowels was tested to see if they obtained response patterns similar to non-ambiguous vowels and also to get an insight into the 'level of processing'. This point will now be more fully discussed below.

It was of interest to see if there were any interactions between consonant responses, vowel responses, and a stimulus characteristic. In particular, the categorization of ambiguous vowels depend upon previously presented syllables (owing to a contrast effect), and thus the same acoustic information could be heard as one vowel at one time, and another vowel at another time (see Thompson and Hollien, 1970). How would this affect the pattern of consonant classifications? Would the pattern of consonant responses to the stimulus parameters (i.e., formant onsets) shift as a function of the label of the ambiguous vowel, or would patterns of response be identical, regardless of how the vowel context was categorized?

The specific F2v values that were chosen in the perceptual study of Chapter 3 provide an extremely good test of the change of patterns of response as a function of vowel label for the following reason. For low F2v's (e.g., the vowels /o/ or /U/), low F2i's were categorized as /b/ or /g/ while high F2i's were categorized as /d/; the *pattern of response* for this case was therefore '[b-g] to /d/'. For high F2v's, on the other hand, low F2i's were categorized as /b/, intermediate F2i values were categorized as /d/ and the highest F2i's were heard as /g/; here, the *pattern of response* was '/b/ to /d/ to /g/'. Thus, as the F2v was raised, the ordinal pattern of consonant categorization responses shifted from a '[b-g] to /d/' pattern to a '/b/ to /d/ to /g/' response pattern. If an ambiguous vowel is labelled as a low F2v vowel, will the consonant response patterns reflect the '[b-g] to /d/' pattern of responses? Similarly, if the same ambiguous vowel is labelled as a vowel with a high F2v, will the '/b/ to /d/ to /g/' response pattern emerge?

If the pattern of consonant responses remained independent of the following vowel label, then subjects could be responding to purely auditory features of the vowel. However, if an interaction was found for consonant categorization, vowel categorization and a stimulus characteristic (e.g., F2i or F3i), subjects could be responding in part to the phonetic label of the vowel.

In this chapter, the *interaction* of several factors are addressed, since the basic effects of F2i and F3i on consonant categorization are generally well established.[1]

1. The interaction of F2i and F3i on the consonant categorizations; Hoffman (1958) and Harris et al. (1958) suggested that F2i and F3i make independent contributions to the categorization of stop consonants. Will the categorization results obtained from the experiment reported in Chapter 3 show F2i and F3i independence?
2. The interaction of the vowel label with consonant categorizations, disregarding specific stimulus characteristics; Mermelstein (1978) tested for this interaction for final consonant categorizations. In his study, the duration of the preceeding vowel was altered and subjects were asked to categorize both the vowel and final consonant. He tested for a Vowel x Consonant interaction and found that, overall, there was no such interaction, but subject differences were apparent. On the other hand, Massaro and Cohen (1983) found a Consonant x Consonant interaction when subject were asked to categorize two consonants in an initial consonant cluster. A CV interaction in the present study could be indicative of a bias towards a particular syllable, or perhaps an indication that the vowel label

---

[1] Quantitative analyses support this assumption; the deletion of either F2i or F3i leads to highly significant changes in the goodness of fit.

has an effect on consonant categorization, irrespective of F2i and/or F3i. Will the results of the perception study of initial consonants done in Chapter 3 show similar results to that of Mermelstein's (i.e., no category by category interactions) or results similar to Massaro and Cohen's (i.e., a significant category by category interaction)?

3. The interaction of the vowel label, consonant categorizations and a stimulus onset characteristic (e.g., F2i or F3i); as mentioned above, this type of interaction would be indicative of an effect of vowel labelling on the *pattern of responses* which is not merely a response bias towards a particular syllable but is rather affected by the acoustic onset characteristics of the stimulus. Massaro and Cohen (1983), in their study done on the perception of initial consonant clusters, showed that a model which had a category by category by stimulus characteristic interaction was not a significant improvement over a model which only had a category by category interaction. Would this study show similar results (i.e., no significant interaction of Consonant x Vowel x Stimulus Characteristics)?

A quantitative method to address the issues of the effects of stimulus variables on categorization and of interdependeicies of category labels is provided by the log-linear approach to multivariate categorical responses with more than two categories. Such methods are discussed by

Fienberg, 1980 (pp. 111-116; see also Haberman, 1979, pp. 369-443). The following discussion of the application of log-linear analysis of data from phonetic categorizaton experiments has been suggested to me by T. Nearey.

It seems reasonable to assume that response profiles to stimuli from a single subject are multinomially distributed and that responses to successive stimuli are independent (see Bock 1975, p. 552). In this case, log-linear methods are directly applicable to the data of a single subject.

There are two types of statistical tests associated with log-linear models. The first of these is the overall goodness of fit test of a single model. The second is the test of the improvement of degree of fit between two 'nested' models (i.e., between a simpler model and a more complex model that includes all the terms of the simpler model.) Unfortunately, no standard significance tests are available for overall goodness of fit when there are a large number of cells with small expected values, as is common in multi-category phonetic experiments. However, tests of the significance of the differences between nested models are still applicable in such cases (see Haberman, 1979, p. 421, or Fienberg, 1980, pp. 174-176 for a discussion of these points.) Therefore, only this latter type of test will be reported here.

Another problem arises in dealing with data from more than one subject, since no simple statistical techniques appear to be available for repeated measures categorical data (Bock, 1975, p. 552), even for the comparison of nested models. While log-linear procedures are strictly applicable only to data from single listeners, it was nonetheless decided to report exploratory analyses on data pooled across speakers for two experiments. This is because the large stimulus set required for investigations of the type discussed in Chapter 3 preclude the possiblility of collecting large numbers of responses to each stimulus from individual subjects. In the analysis of such pooled data, the ordinary tests of significance associated with log-linear models are valid only under the strong and unlikely assumption of subject homogeneity (i.e. that there is no difference between subjects in their responses to any of the stimuli.) In no case are any substantative conclusions based solely on such nominally significant results. Rather such statistics are used to give some indication of the relative magnitude of certain effects, to compare with graphic analyses and to concentrate attention on specific issues for further experiments.

Some theoretically important issues exhibiting nominally signficant reusults are investigated in a final experiment involving a subrange of stimuli. In this experiment, a large number of responses is collected from each subject for each stimulus and analyses are carried out on a subject by subject basis. In this last experiment, ordinary tests for differences between models can be applied with no known violations of assumptions.

### 4.1 22 Subjects Pooled

#### 4.1.1 Procedure

The subjects, stimuli and procedure were identical to those in Chap. 3.

#### 4.1.2 Results and Analysis

The two data sets with 'ambiguous' vowels were [o-U] and [U-ɛ]. An analysis of the specific interactions of interest was done on the two data sets for 22 subjects pooled.

Table 4.1 shows the results of the log-linear analysis for the 'ambiguous' vowel [o-U]. Three models were compared, corresponding to the three questions of interest.

First of all, F2i and F3i appear to make independent contributions to the consonant categorizations, since a model having a F2i x F3i x Consonant interaction does not significantly improve the fit goodness of fit to the data as compared to a model in which only two-way interactions are included. This result concurs with Hoffman's (1958) and Harris et al.'s (1958) suggestion of F2 and F3 independence.

Table 4.1 also shows that there is a significant improvement of the goodness of fit if the model includes a Consonant x Vowel interaction; this suggests the possibility that there are response biases to particular syllables. This result does not corroborate Mermelstein's (1978) result of no overall significant category by category interaction for

## TABLE 4.1

## Log-Linear Analyses; 22 Subjects Pooled; Vowel [o-U]

## (F2v=1050 Hz)

| Model | D.F. | $G^2$ |
|---|---|---|
| 1. CV, FG, FC, GC | 368 | 294.19 |
| 2. CV, FG, FC, GC, FV | 360 | 287.08 |
| 3. CV, FG, FC, GC, FV, GV | 352 | 277.12 |
| 4. CV, FGC | 240 | 156.69 |
| 5. C, V, FG, FC, GC | 370 | 458.09 |
| 6. FG, FCV,GCV | 320 | 243.16 |

| Comparison of Models | D.F. | $G^2$ |
|---|---|---|
| Model 1 vs. Model 4 | 128 | $137.5^{n.s.}$ |
| Model 1 vs. Model 5 | 2 | $163.9^{**}$ |
| Model 3 vs. Model 6 | 32 | $33.96^{n.s.}$ |

Key: C= Consonant /b/, /d/ or /g/
V= Vowel /o/ or /U/
F= F2i
G= F3i

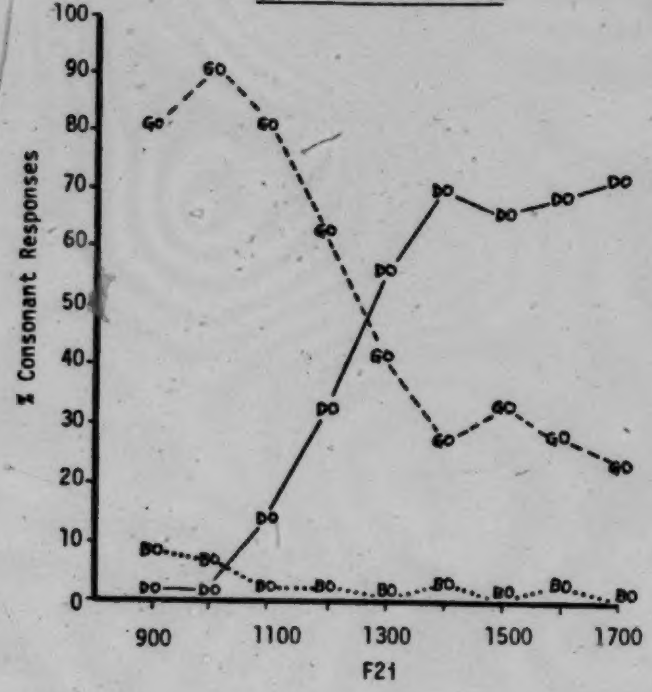
Three-way interactions include all possible two-way interactions.
n.s. indicates non-significance
** indicates significance to the .01 level

final consonant categorization, but recall that he also found subject differences. However, Massaro and Cohen (1983) did find such category by category interactions in initial consonant clusters.

Finally, a model with a three-way interaction between the consonant response, the vowel response and a stimulus characteristic did not siginificantly improve the goodness of fit, as compared to a model without this three-way interaction. As seen in Table 4.1, the addition of this factor was not significant; thus, the patterns of responses do not seem to be affected by the vowel label. This result is similar to that of Massaro and Cohen's (1983).

Figure 4.1 shows a plot of the marginal values of consonant classifications for F2i (pooled over F3i), depending on the vowel response, /o/ (top plot) or /U/ (bottom plot). Percent consonant responses were calculated for each F2i value, given the vowel label. The patterns of response in the two plots of Figure 4.1 look similar; subjects respond with a majority of /g/'s at lower F2i levels, and respond with a majority of /d/'s at higher F2i levels. The graphs show that low F2i's yields more /b/ responses when subjects categorize the 'ambiguous' vowel as being /U/, than when they categorize the vowel as /o/. However, the level of /b/ responses in both cases is fairly low (below 35% of the total number of consonant responses when the vowel is labelled /U/, and below 10% of the total number of consonant responses when the vowel is labelled

Vowel Labelled As /o/



Vowel Labelied As /U/

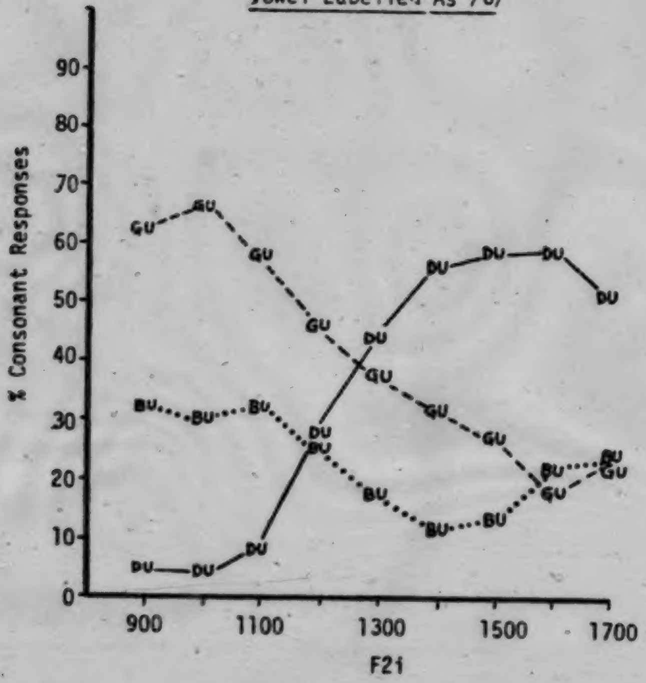


FIGURE 4.1 F2i Marginals; Ambiguous Vowel [o-U]: 22 Subjects Pooled

/o/).

In general, the patterns of response were similar whether the vowel was labelled /U/ or /o/. When subjects label the vowel as being /o/, majority responses for /d/ and /g/ are at a higher level than when the vowel is labelled as /U/, but the specific majority responses are the same under both labelling conditions. Responses for /bU/ are higher than for /bo/, but there are only a small number of total /b/ responses. Thus, subjects could be operating at an 'acoustic level' of processing, since phonetic labelling of the vowel did not significantly alter their consonant responses.

Table 4.2 shows the results of a log-linear analysis for the 'ambiguous' vowel [U-ɛ]. Three comparisions of models were made to test the specific questions of interest.

Again, F2i and F3i seem to be independent, as suggested by Hoffman (1958) and Harris et al. (1958), as the addition of an F2i x F3i interaction did not significantly improve the goodness of fit to the data.

A model which included a Consonant x Vowel interaction significantly improved the goodness of fit as compared to a model without this interaction. This perhaps indicates a response bias to a particular syllable, which, again, does not corroborate Mermelstein's (1978) results but is similar to Massaro and Cohen's (1983) findings.

A model which included a Consonant x Vowel x Stimulus Characteristic (i.e., F2i or F3i) in this case did

## TABLE 4.2
## Log-Linear Analyses; 22 Subjects Pooled; Vowel [U-ɛ]

(F2v=1450 Hz)

| Model | D.F. | $G^2$ |
|---|---|---|
| 1. CV, FG, FC, GC | 216 | 226.29 |
| 2. CV, FG, FC, GC, FV | 210 | 225.07 |
| 3. CV, FG, FC, GC, FV, GV | 204 | 224.28 |
| 4. CV, FGC | 144 | 135.35 |
| 5. C, V, FG, FC, GC | 218 | 258.57 |
| 6. FG, FCV, GCV | 180 | 181.90 |

| Comparison of Models | D.F. | $G^2$ |
|---|---|---|
| Model 1 vs. Model 4 | 72 | $90.94^{n.s.}$ |
| Model 1 vs. Model 5 | 2 | $32.28^{**}$ |
| Model 3 vs. Model 6 | 24 | $42.38^{**}$ |

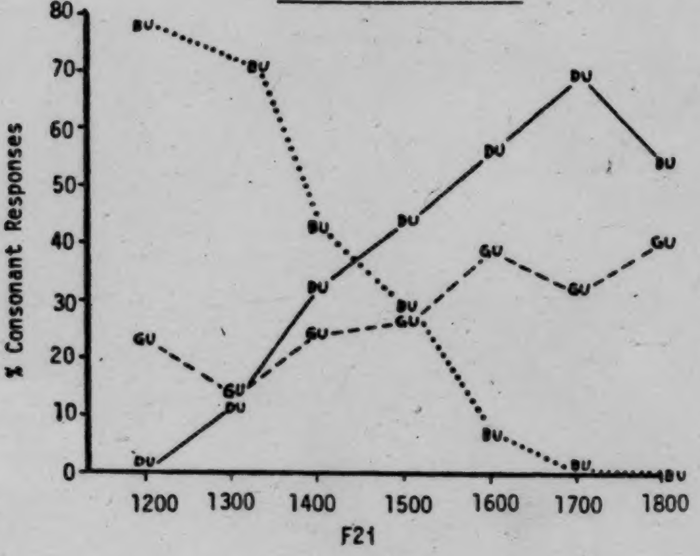
Key: C= Consonant /b/, /d/ or /g/
V= Vowel /U/ or /ɛ/
F= F2i
G= F3i

Three-way interactions include all possible two-way interactions.
n.s. indicates non-significance
** indicates significance to the .01 level

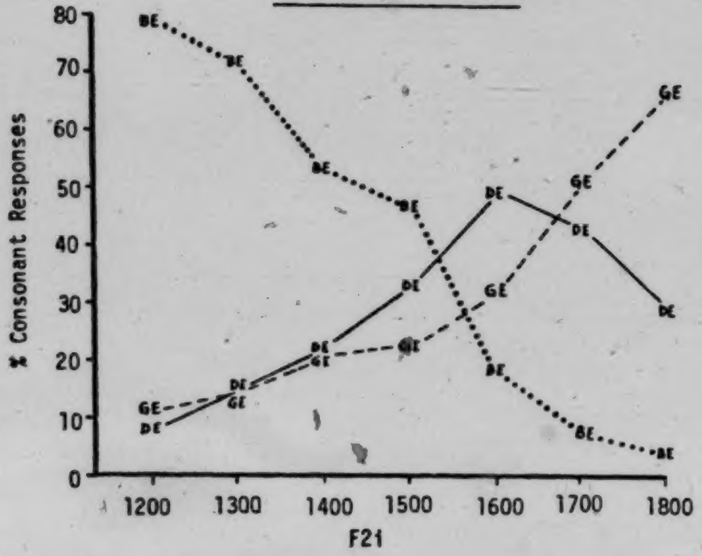
significantly improve the goodness of fit to the data, contrary to the case of the vowel [o-U], and also contrary to Massaro and Cohen's (1983) results.

Figure 4.2 shows a plot of the marginal values of consonant classifications for F2i (pooled over F3i), depending on the vowel response, /U/ (top plot) or /ɛ/ (bottom plot). The graphs show that majority /b/ responses are somewhat higher when the vowel is labelled /ɛ/ than when labelled as /U/ but the pattern of response for /b/ categorizations are similar. It is interesting to note that for the ambiguous [o-U] vowel, /b/ responses were higher when the vowel was labelled /U/ than when it was labelled /o/; here, /b/ responses are higher when an ambiguous vowel is labelled /ɛ/ than when it is labelled /U/. In both cases, the preferred vowel is the one which has a higher F2v in natural data; this could be interpretted as an interesting case of what has been referred to as the 'overshoot' phenomenon in vowel perception (see Lindblom and Studdert-Kennedy, 1967).

On the other hand, for /d/ and /g/ categorizations, vowel labelling as a function of formant onsets seems to make a difference as to how the subject categorizes the consonant. At the high levels of F2i (e.g., F2i=1700 Hz or 1800 Hz), the majority stop consonant response is /g/ when subjects label the ambiguous vowel as being /ɛ/, but is /d/ when subjects label the vowel as being /U/.



Vowel Labelled As /U/



Vowel Labelled As /ɛ/

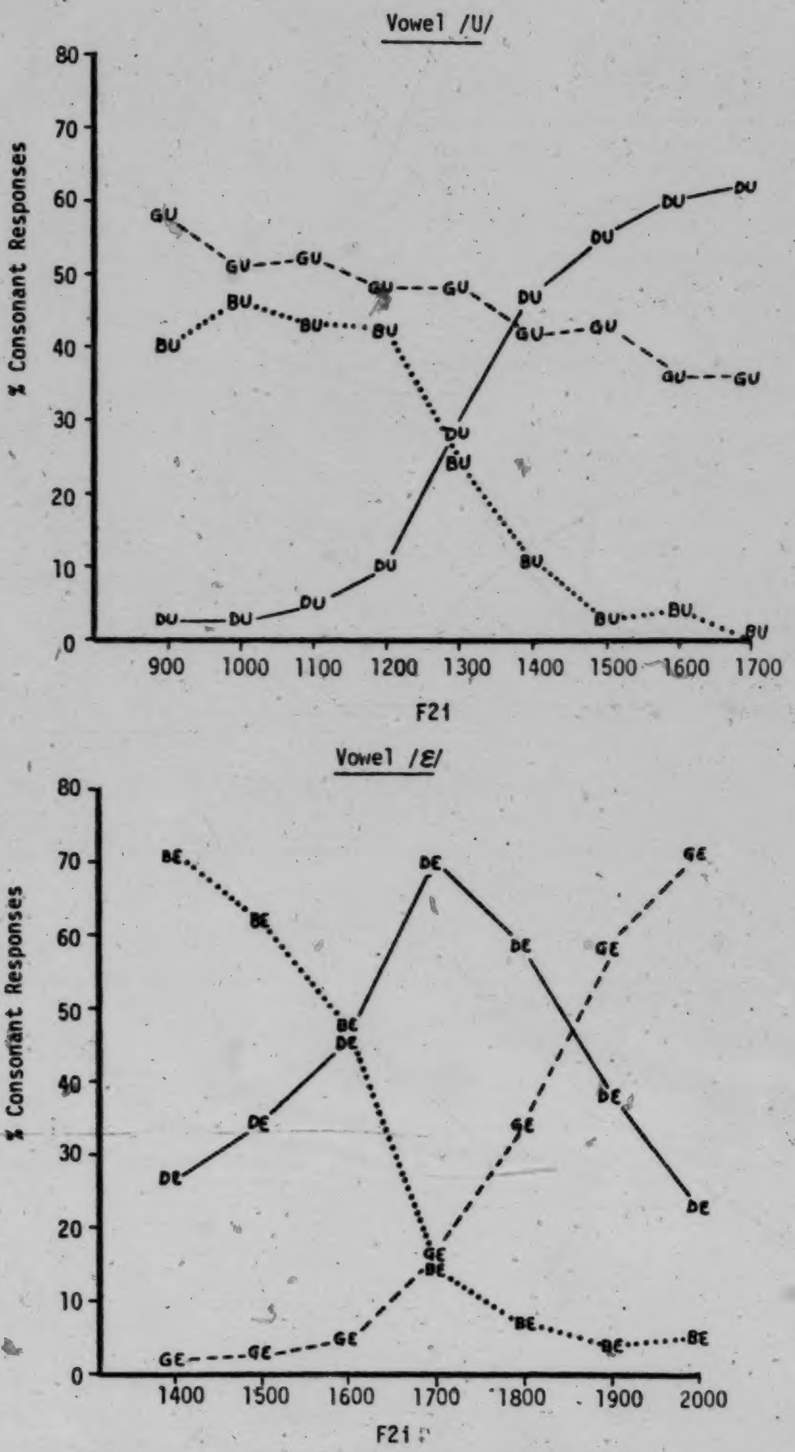
FIGURE 4.2 F2i Marginals; Ambiguous Vowel [U-ɛ]: 22 Subjects Pooled

Figure 4.3 shows the plot of the F2i marginals (pooled over F3i) for the responses from the /U/ (F2v=1050 Hz) vowel set (top plot) and for the responses from the /ɛ/ (F2v=1650 Hz) vowel set (bottom plot). Comparing Figures 4.2 and 4.3 it can be seen that the patterns of response for ambiguous [U-ɛ] look similar to patterns obtained in the /ɛ/ set *when subjects label the ambiguous vowel as being /ɛ/*. On the other hand, when the ambiguous vowel is labeled /U/, the patterns of response are different than when the vowel is labelled as /ɛ/, but are also dissimilar in one respsect to the patterns of response from the /U/ vowel set. In particular, there is a majority of /b/ responses in the the ambiguous vowel set when the ambiguous vowel is labelled as /U/, but the response pattern in the /U/ vowel set shows no majority /b/ responses. However, the /d/ and /g/ response patterns of the /U/-labelled ambiguous vowel look quite similar to the /d/ and /g/ response patterns from the /U/ vowel set.

#### 4.1.2.1 Analysis of the Effect of the Phonetic Label

Were subjects responding on the basis of the phonetic label of the vowel? Did the categorization of the vowel in some sense determine the consonant categorizations? One way to examine the role of vowel labelling is to define 'syllable templates' (cf. Lieberman, 1979).

One possible definition is that a syllabic template is the average value of onsets and formant tracks of



Vowel /U/

Vowel /ɛ/

FIGURE 4.3 F2i Marginals; Non-ambiguous Vowels /U/ and /ɛ/: 22 Subjects Pooled

measured syllables. Such syllabic templates can be easily obtained from the measurement study of Chapter 2. Examples of the F2 and F3 portion of /bɛ/, /bU/, /dɛ/, /dU/, /gɛ/ and /gU/ syllables are shown in Figure 4.4, which are based on the average male onset and steady state values measured in Chapter 2.

Syllabic templates evaluated 'simultaneously' would classify an incoming signal as being a particular syllable if the parameters (in this case, formant parameters) of this signal were closer to the parameters of this syllable's stored template than to any other stored template. All parts of the template would be compared 'simultaneously' to the incoming signal. A 'hierarchical' evaluation of syllabic templates would first categorize the vowel portion and then match the formant onsets of the incoming signal to the closest template formant onset for that vowel. The latter approach will be used in this analysis, since the vowel portion is 'ambiguous' and thus would presumably not count in the distance from template measure, since the vowel portions would be roughly the 'same distance' away from the two templates having the non-ambiguous vowels (e.g., 'ambiguous' [U-ɛ] would be as far away from 'good' /U/ as from 'good' /ɛ/). The F2 portion of the templates is of most interest since the vowel label is mainly a function of F2v. The F2 onset values of the synthetic stimuli are plotted on the left hand side of

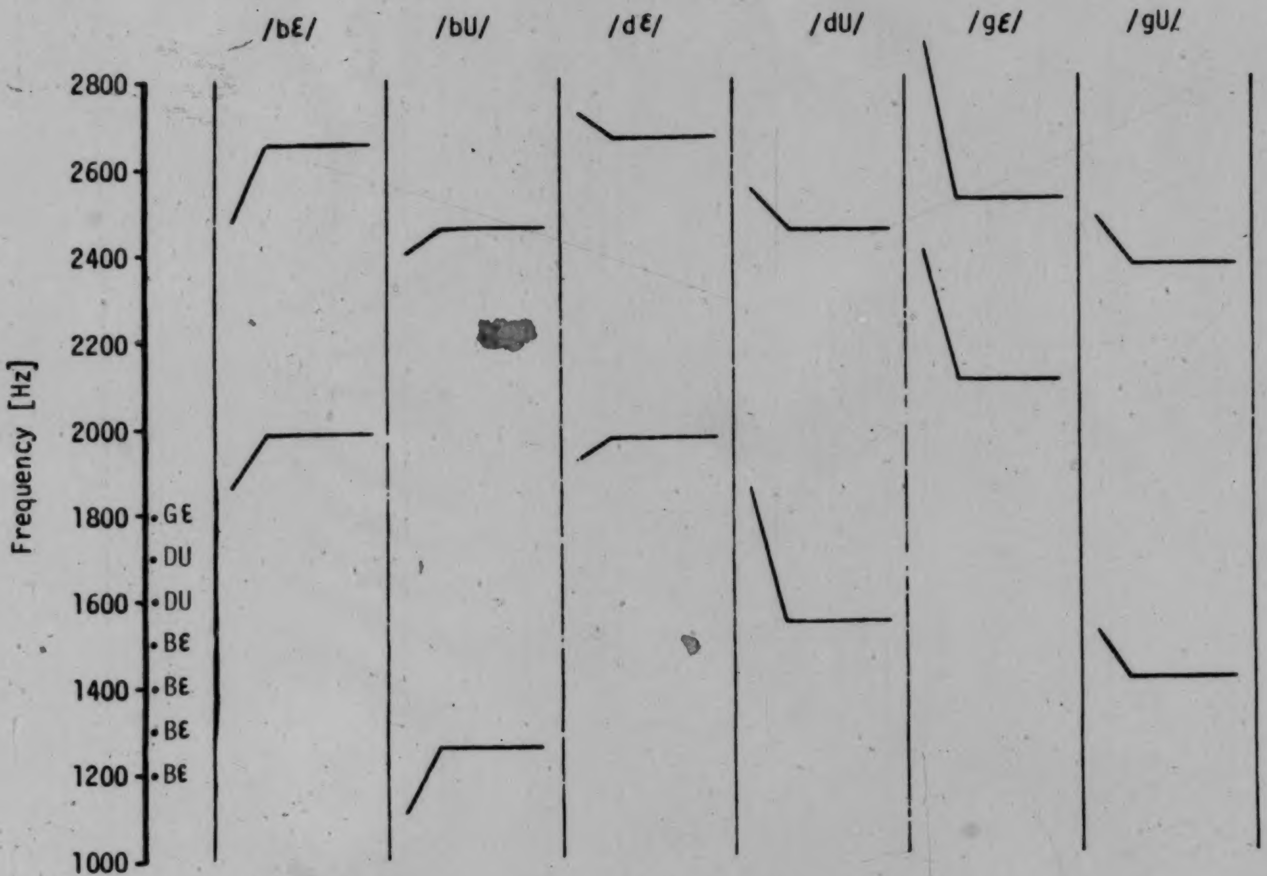


FIGURE 4.4 Syllabic Templates From Average Male Formant Measurements

Figure 4.4 and are labelled with the syllable which obtained the highest number of responses.

Do the categorizations reflect 'value closest to the appropriate onset of a template'? For synthetic stimuli having F2i=1200 Hz and F2i=1300 Hz, the majority response was /bɛ/; in fact, these values lie much close to the /bU/ template. For the synthetic stimuli with F2i=1400 Hz, and F2i=1500 Hz, majority response was /bɛ/ but these onset values are closest to template /gU/. For the synthetic stimuli with F2i=1600 Hz and F2i=1700 Hz, the majority response was /dU/. In fact, F2i=1600 Hz is closest to template /gU/, while F2i=1700 Hz is closest to template /dU/. The synthetic stimuli with F2i=1800 Hz obtained majority /gɛ/ responses, but this onset value lies closest to the /dU/ template. [7]

Thus, though there is some evidence of a vowel labelling factor, it cannot be easily explained in terms of the onsets of a formant based syllable-template theory.

In this study, responses were pooled over subjects. Would the overall pattern of responses be similar when subjects were tested with more randomizations? In this experiment (with 22 subjects pooled) the 'ambiguity' of the vowel was really an indication that half of the subjects heard the 'ambiguous' vowel as a particular vowel, while the other half heard it as another vowel. In the following experiment, replications by eight subjects were done, so that the 'ambiguity' of the vowel also entails intra-subject vowel labelling factors (i.e., the subject 'hears' one vowel one time and another vowel another time owing to contrast effects).

---

[7] Preliminary anlalysis indicates that similar results are obtained when both the onsets and steady-states of the stimuli are compared to the onset and steady-state values of the templates.

## 4.2 Eight Subjects, Five Replications

### 4.2.1 Speakers

Eight graduate linguistic students, with no history of speech or hearing disorders, were used. All subjects were native Canadian English speakers.

### 4.2.2 Stimuli and Procedure

The stimuli and procedure were identical to those described in the Main Perception Experiment (F2v Experiment) in Chapter 2. However, five randomizations of the stimulus set were presented to each subject, one randomization at each of five separate sittings.

### 4.2.3 Results and Analysis

Results of each subject's consonant categorizations are given in Figures 4.5-4.12. Indications of subject differences are evident, though it should be noted that only tentative conclusions can be reached since the number of data points for each stimulus item is small (only five

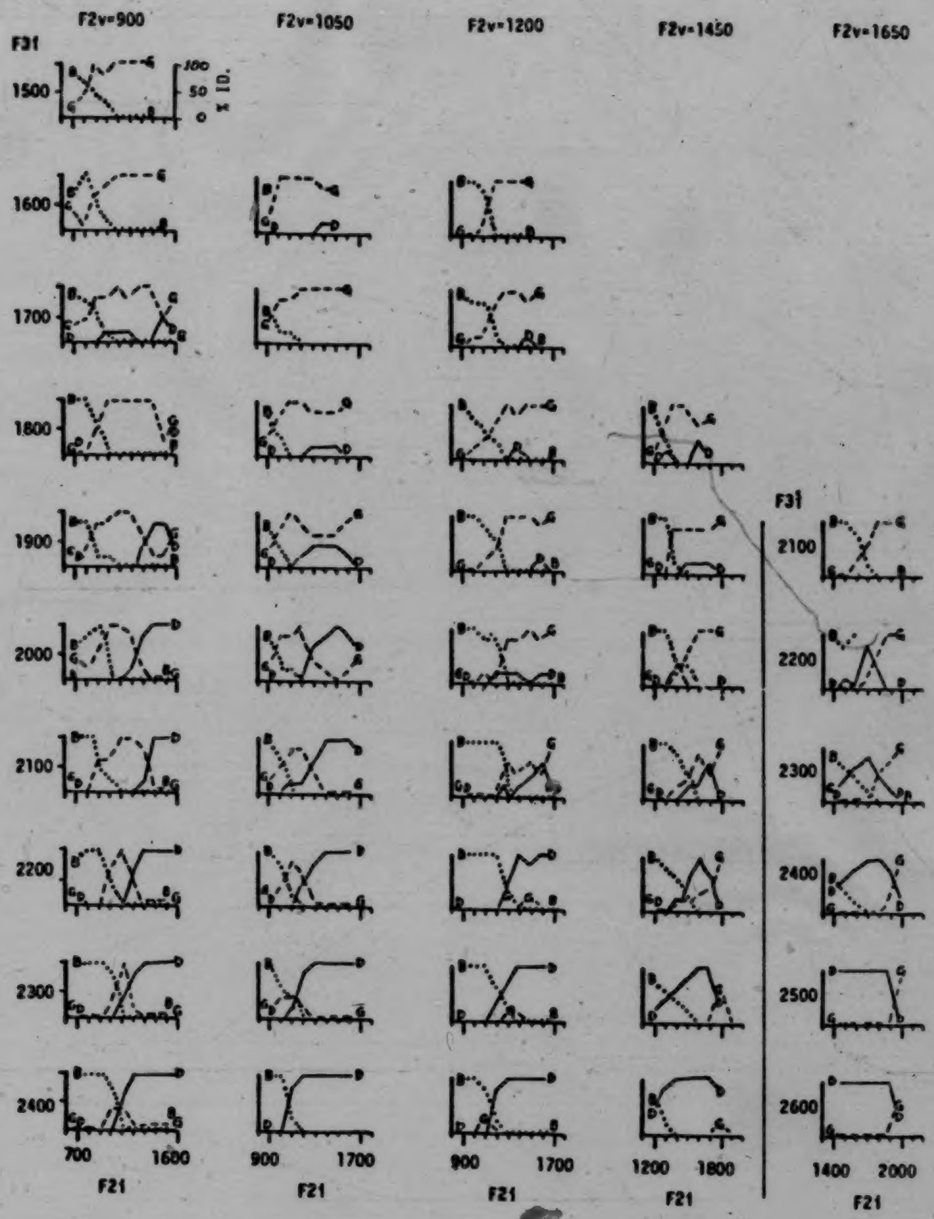


FIGURE 4.5 Consonant Categorizations; F2v Experiment; Subject SS

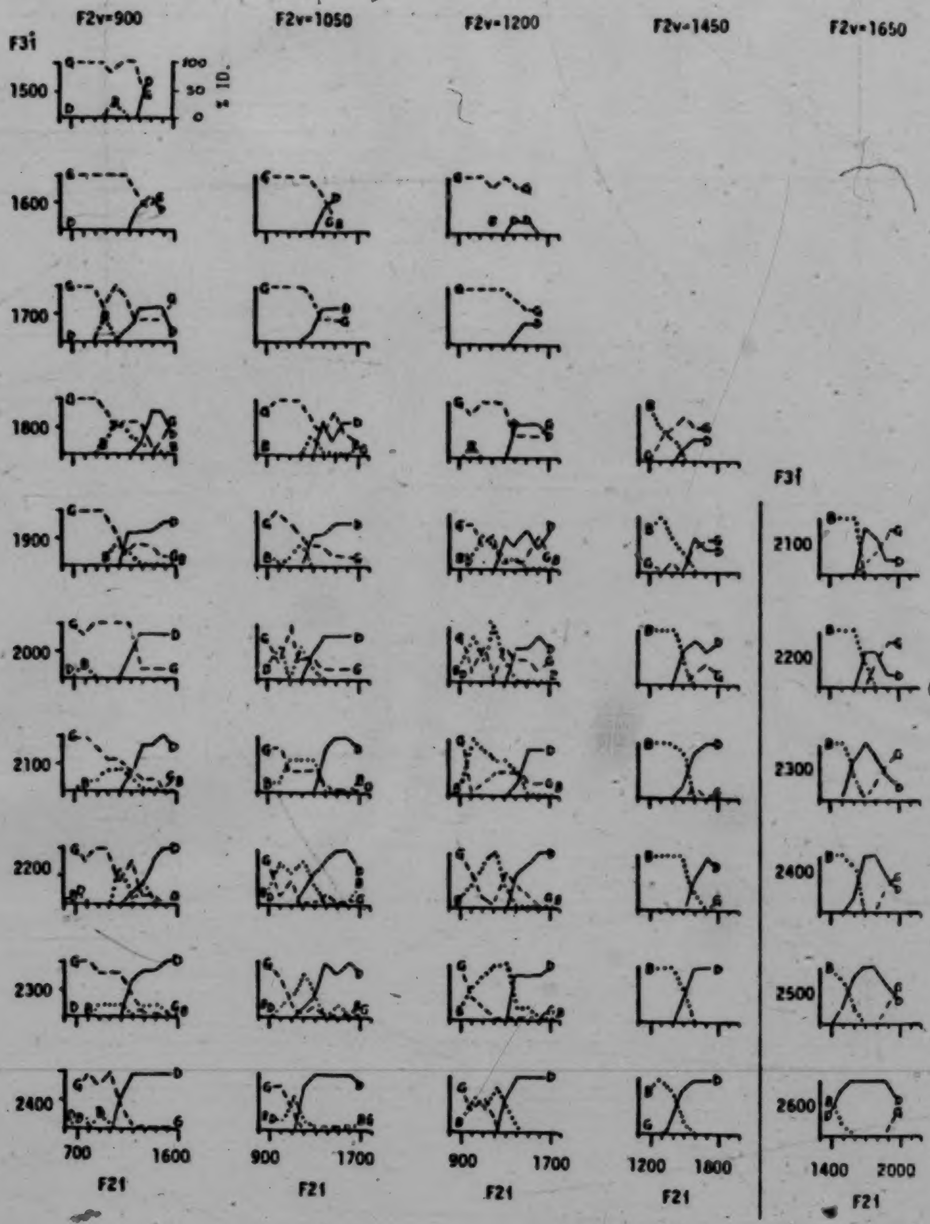


FIGURE 4.6 Consonant Categorizations; F2v Experiment; Subject TR

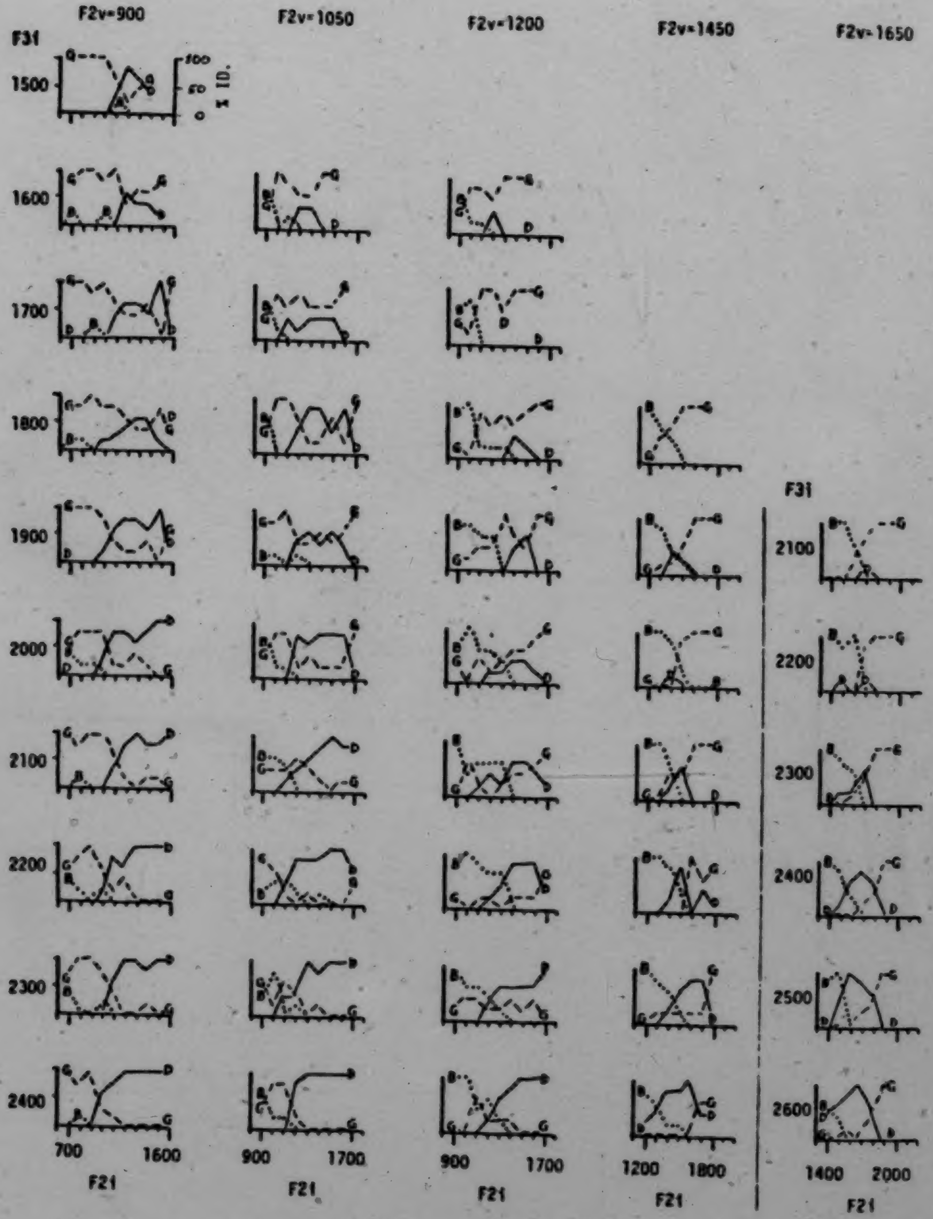


FIGURE 4.7 Consonant Categorizations; F2v Experiment; Subject PA

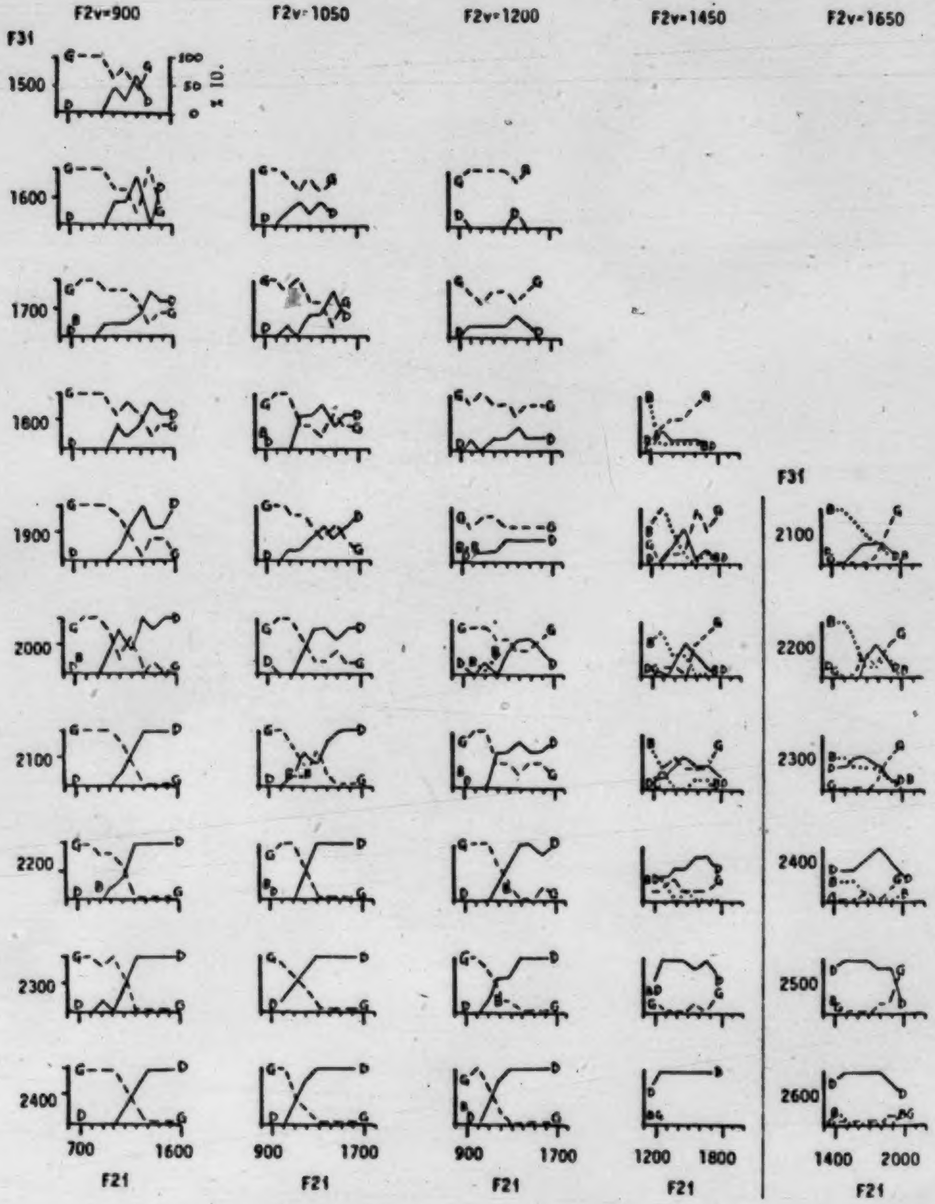


FIGURE 4.8 Consonant Categorizations; F2v Experiment; Subject BC

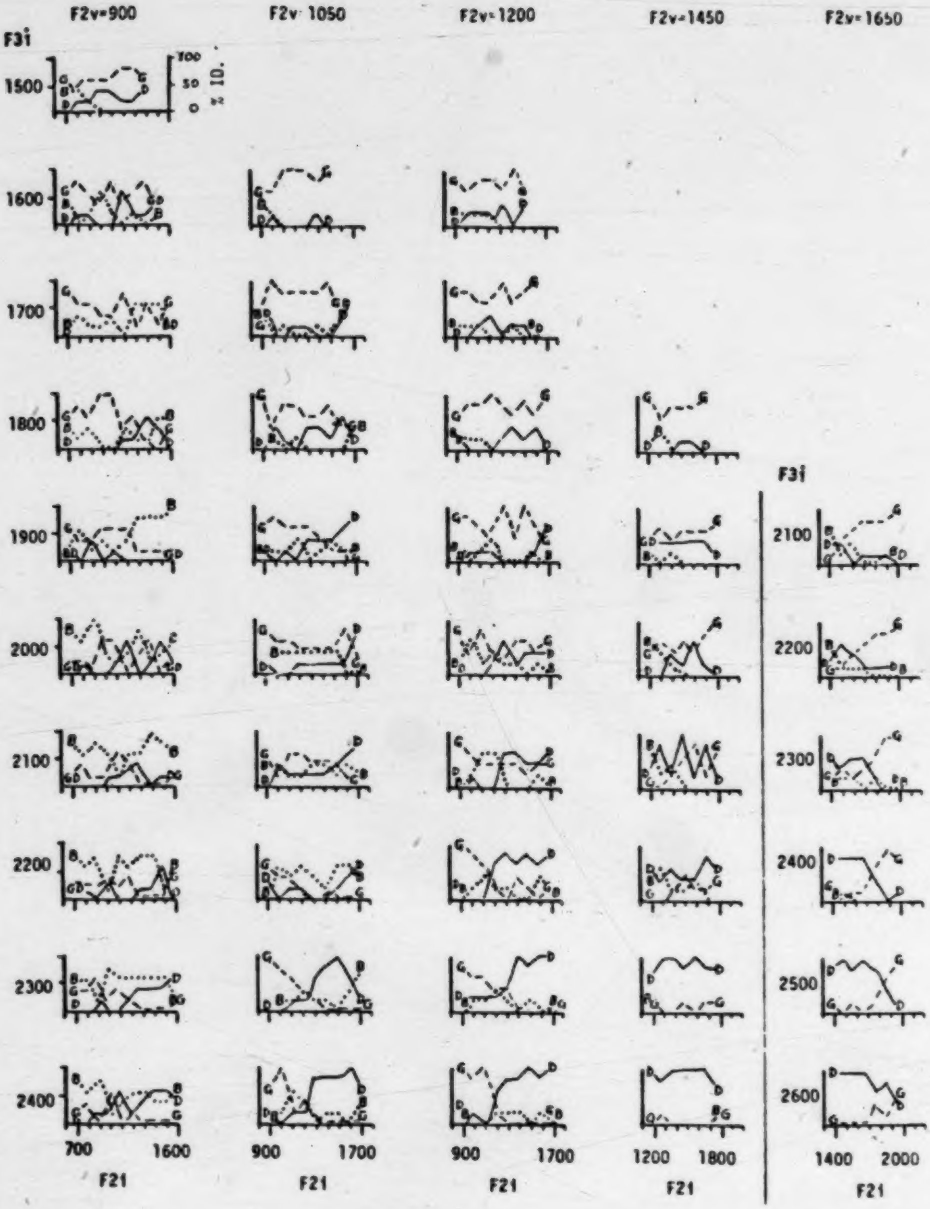


FIGURE 4.9 Consonant Categorizations; F2v Experiment; Subject RH

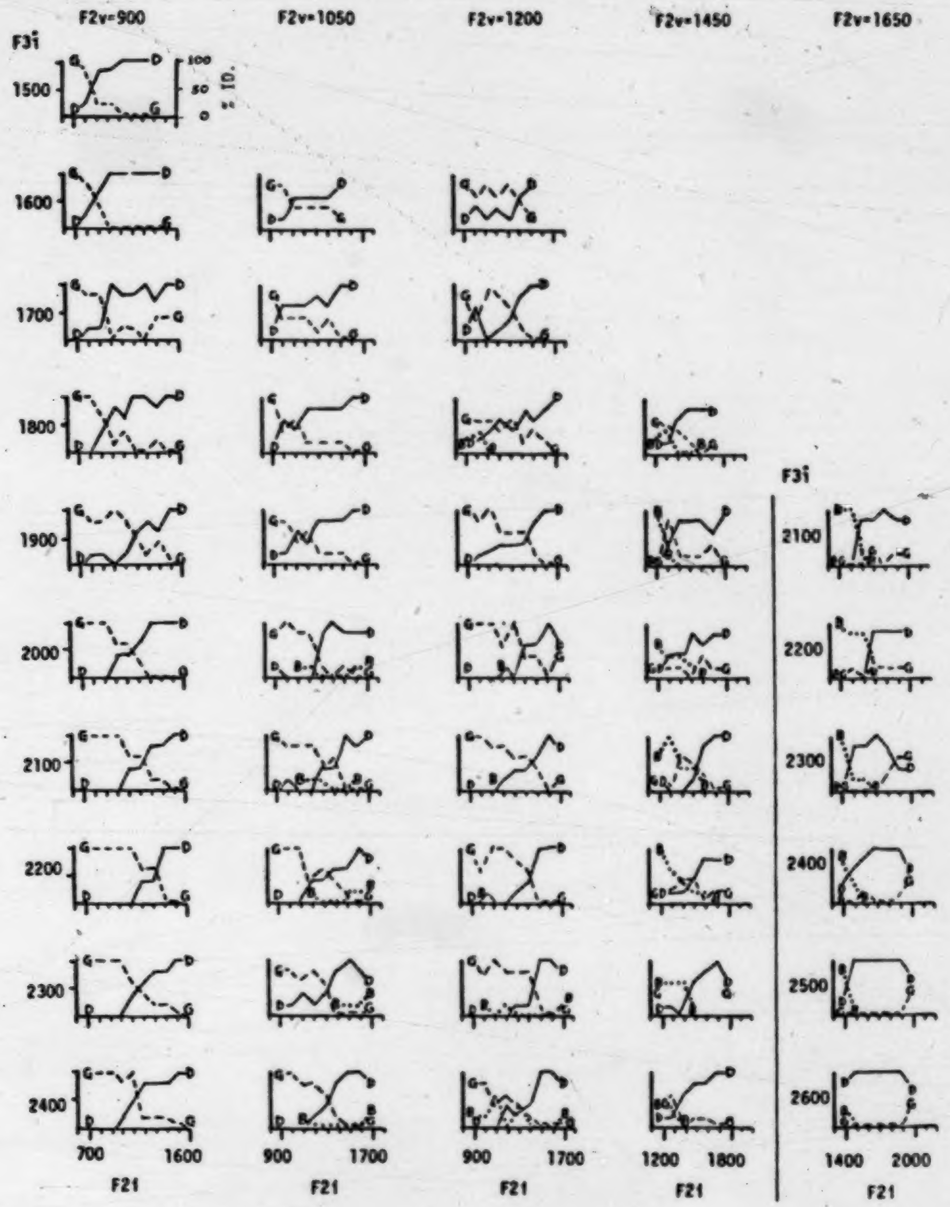


FIGURE 4.10 Consonant Categorizations; F2v Experiment; Subject MM

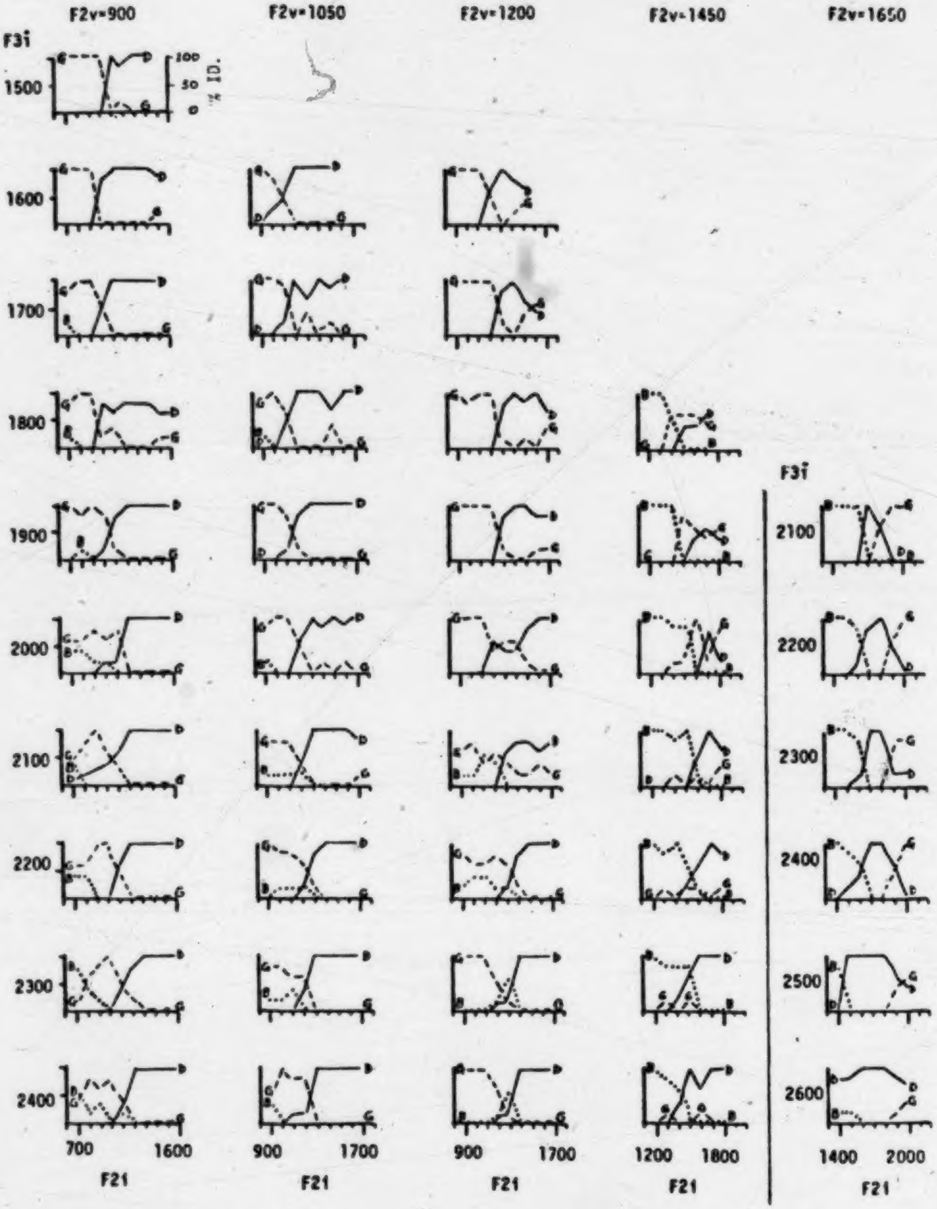


FIGURE 4.11 Consonant Categorizations; F2v Experiment; Subject TD

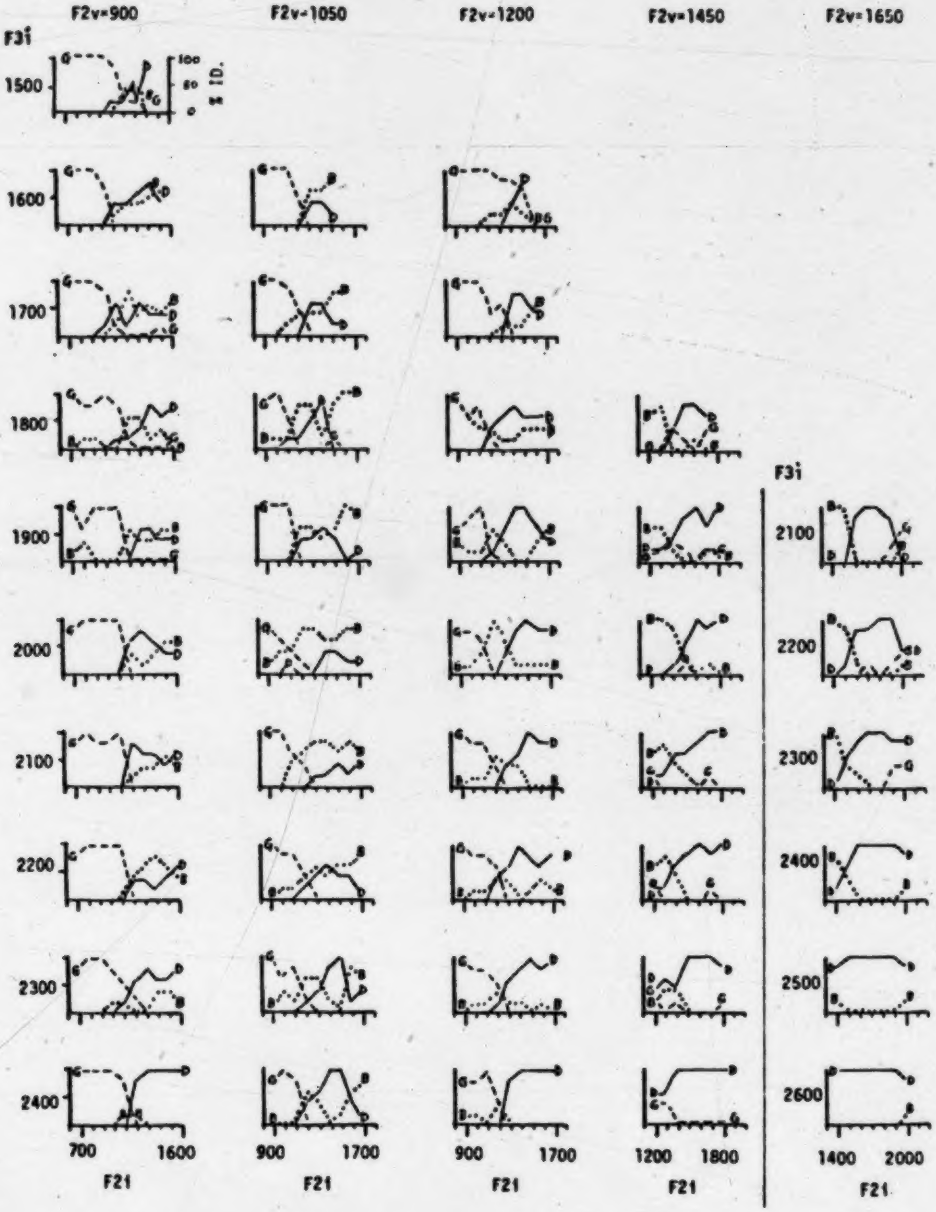


FIGURE 4.12 Consonant Categorizations; F2v Experiment; Subject JS

randomizations were done).

Subjects SS, TR, PA, BC and RH (Figures 4.5-4.9, respectively) have rather similar consonant cetegorizations since, for these subjects, changes in the F3i produce consonant categorization changes. Subjects MM, TD and JS (Figures 4.10-4.12, respectively), on the other hand, show less consonant category change as F3i changes, and seem to respond more to F2i changes. Thus, there seems to be subject differences with regard to the strength of F3i cues.

In general, the patterns of response shown by the large group of subjects, pooled together in the main experiment, are also evident in these results. F2v, F3i and F3i all contribute to consonant categorization, though, as mentioned above, F3i seems to be a weaker cue for some subjects than for others. It should be noted that other investigators have also found similar subject differences (e.g., Walley and Carol, 1983).

Did vowel categorization affect consonant categorization? Owing to the small number of replications, this question was not addressed on a subject-by-subject basis. However, pooling all eight subjects with five randomizations each (giving 40 data points per cell), log-linear analyses were done, as in the previous section. Although this pooling procedure constitutes a violation of the log-linear requirements, nevertheless, it was felt that at least an indication of interaction patterns could be obtained. Thus, though these results are not definitive by any means, they may show general trends of responses which could then be more fully investigated.

The results of a log-linear analysis for the 'ambiguous' vowel [o-U] are shown in Table 4.3. These show that the addition of a F2i x F3i interaction is significant, which does not support the idea of F2i and F3i independence proposed by Hoffman (1958) and Harris et al. (1958). Preliminary analysis indicates that stimuli having high F2i's and low F3i's deviate considerably from values which are predicted by a model of independence. These are precisely the stimuli which would have a 'compact' spectral onset shape (see analysis of spectral shape in Chapter 3).[*]

Table 4.3 also shows that the Consonant x Vowel interaction significantly improves the goodness of fit; this is again perhaps an indication of a response bias to a particular syllable contrary to Mermelstein's (1978) overall result of no significant CV interaction, but which corroborate Massaro and Cohen's (1983) result of a Consonant x Consonant interaction in the perception of initial consonant clusters.

Finally, results show that addition of a three-way interaction of consonant response, vowel response and stimulus characteristic is non-significant, which is similar

---

[*] Preliminary analysis of the other vowel sets also indicates that those having low F2v's are more adequately modeled with F2i and F3i interactions, while those having high F2v's are adequately modeled by independent F2i and F3i's. Cell analysis of low F2v sets indicates that stimuli with high F2i's and low F3i's (those with 'compact' shapes) deviate considerably from those predicted by models of independent F2i and F3i's.

## TABLE 4.3
## Log-Linear Analyses; 8 Subjects x 5 Replications (Pooled)
### Vowel [o-U] (F2v=1050 Hz)

| Model | D.F. | $G^2$ |
|---|---|---|
| 1. CV, FG, FC, GC | 368 | 536.50 |
| 2. CV, FG, FC, GC, FV | 360 | 527.03 |
| 3. CV, FG, FC, GC, FV, GV | 352 | 519.74 |
| 4. CV, FGC | 240 | 277.09 |
| 5. C, V, FG, FC, GC | 370 | 776.61 |
| 6. FG, FCV, GCV | 320 | 504.68 |

| Comparison of Models | D.F. | $G^2$ |
|---|---|---|
| Model 1 vs. Model 4 | 128 | 259.41** |
| Model 1 vs. Model 5 | 2 | 240.11** |
| Model 3 vs. Model 6 | 32 | 15.06n.s. |

Key: C= Consonant /b/, /d/ or /g/

V= Vowel /o/ or /U/

F= F2i

G= F3i

Three-way interactions include all possible two-way interactions.
n.s. indicates non-significance
** indicates significance to the .01 level

to Massaro and Cohen's (1983) results.

F2i marginal values for the ambiguous [o-U] vowel are shown in Figure 4.13. The response patterns look similar regardless of the label of the ambiguous vowel. /b/ responses are higher when the vowel is labelled /U/ than when it is labelled /o/, but there are few /b/'s in general.

Table 4.4 shows the results of the log-linear analysis of the [U-ɛ] vowel set. These are

1. the addition of an F2i x F3i interaction does not significantly improve the goodness of fit, which supports the idea of F2 and F3 indepedence;
2. the addition of a Consonant x Vowel interaction significantly improved the goodness of fit to the data (contrary to results of Mermelstein, 1978, but corroborative of Massaro and Cohen's results, 1983), perhaps indicating a response bias to a particular syllable; and
3. the addition of three-way interaction between Consonant x Vowel x Stimulus Characteristic was not significant indicating that the patterns of consonant responses were not significantly different depending on the vowel label (which is similar to Massaro and Cohen's findings, 1983).

Marginal values of responses of F2i pooled over F3i for the ambiguous [U-ɛ] vowel are shown in Figure 4.14. Majority responses are similar, regardless of whether the vowel is labelled as /U/ or /ɛ/, though there are fewer /d/'s in the

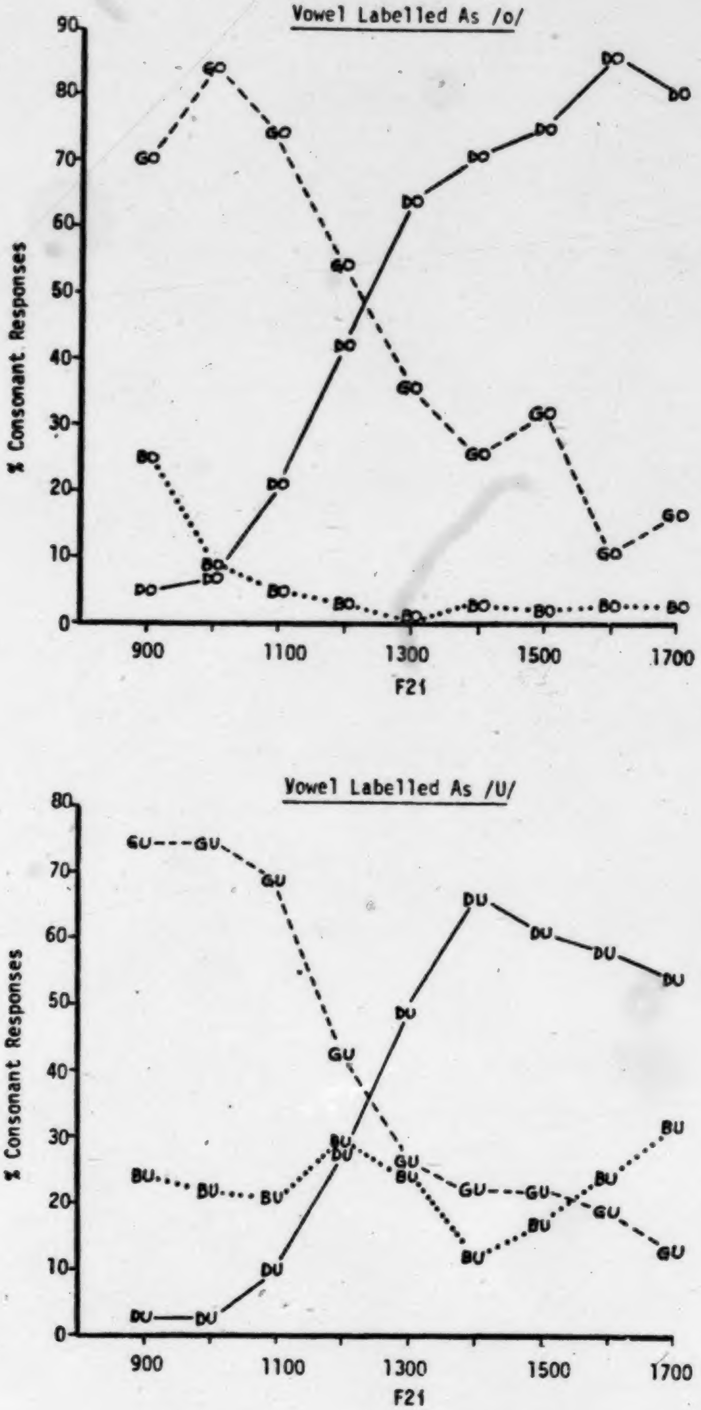


FIGURE 4.13 F2i Marginals; Ambiguous Vowel [o-U]; 8 Subjects x 5 Repetitions

## TABLE 4.4

### Log-Linear Analyses; 8 Subjects x 5 Replications (Pooled)

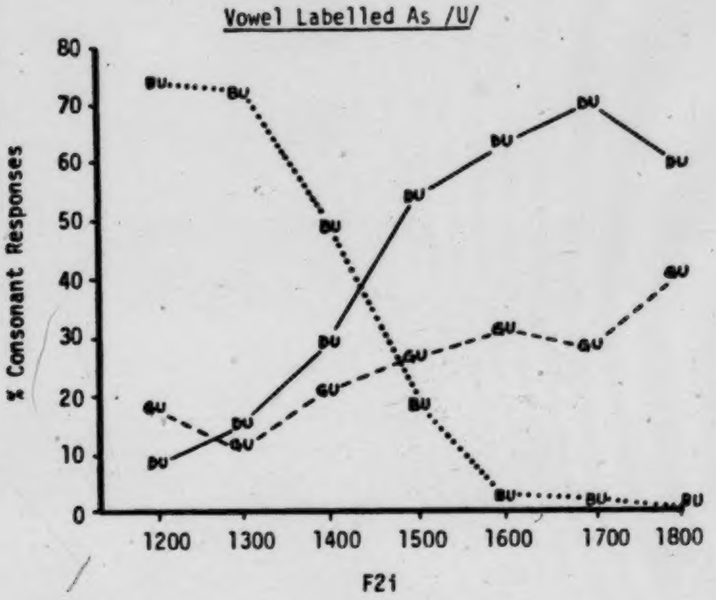
### Vowel [U-ɛ] (F2v=1450 Hz)

| | Model | D.F. | $G^2$ |
|---|---|---|---|
| 1. | CV, FG, FC, GC | 216 | 318.59 |
| 2. | CV, FG, FC, GC, FV | 210 | 259.44 |
| 3. | CV, FG, FC, GC, FV, GV | 204 | 241.25 |
| 4. | CV, FGC | 144 | 227.46 |
| 5. | C, V, FG, FC, GC | 218 | 335.73 |
| 6. | FG, FCV, GCV | 180 | 208.68 |

| Comparison of Models | D.F. | $G^2$ |
|---|---|---|
| Model 1 vs. Model 4 | 72 | $91.13^{n.s.}$ |
| Model 1 vs. Model 5 | 2 | $17.14^{**}$ |
| Model 3 vs. Model 6 | 24 | $32.57^{n.s.}$ |

Key: C= Consonant /b/, /d/ or /g/
V= Vowel /U/ or /ɛ/
F= F2i
G= F3i

Three-way interactions include all possible two-way interactions.
n.s. indicates non-significance
** indicates significance to the .01 level



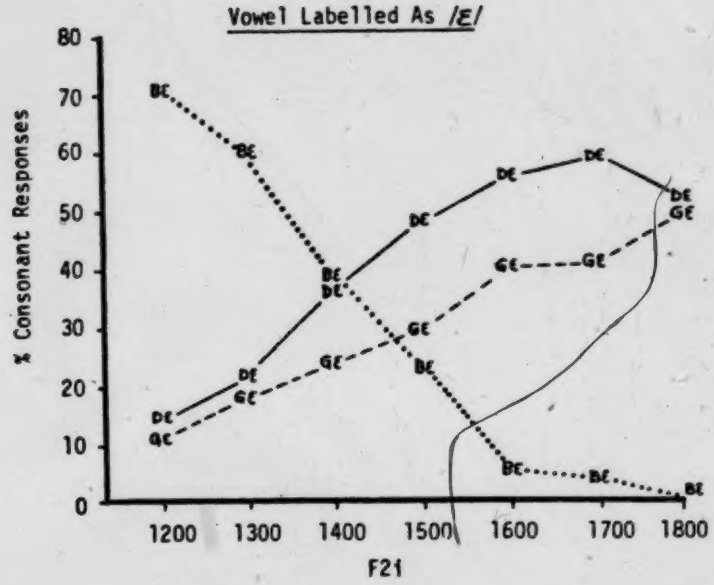


FIGURE 4.14 F2i Marginals; Ambiguous Vowel [U-ɛ]; 8 Subjects x 5 Repetitions

/ɛ/ case than when the vowel is labelled /U/.

Figure 4.15 shows the F2i marginals for the /U/ vowel set (top plot) and the /ɛ/ vowel set (bottom plot) for eight subjects with five repetitions. These plots show response patterns similar to those obtained for /U/ and /ɛ/ with 22 subjects pooled (Figure 4.3).

A comparison of Figure 4.14 (ambiguous [U-ɛ]) with Figure 4.15 (vowel /U/ and vowel /ɛ/) shows that consonant responses when the ambiguous vowel is labelled /ɛ/ *do not* show similar response patterns to those obtained in the /ɛ/ vowel set, and consonant responses when the ambiguous vowel is labelled /U/ *also do not* show similar reponse patterns to those obtaned in the /U/ vowel set. Note that the patterns of consonant responses are fairly similar (i.e., show the same trends) regardless of vowel categorization, unlike the response patterns obtained with 22 pooled subjects. However, response patterns of this subject group were similar to those of the 22 subjects pooled for the /U/ and /ɛ/ vowel sets. In other words, the response patterns of the 'good' vowel sets were replicated, but not the response patterns of this 'ambiguous' vowel set. Though some violations of the log-linear analysis occurred (because of the pooling process), it is interesting to note that, with the group of 22 subjects, vowel categorization affected the pattern of consonant categorization, while with this subgroup of eight subjects, no such interaction was found. It is felt that if this interaction was a very strong factor in consonant

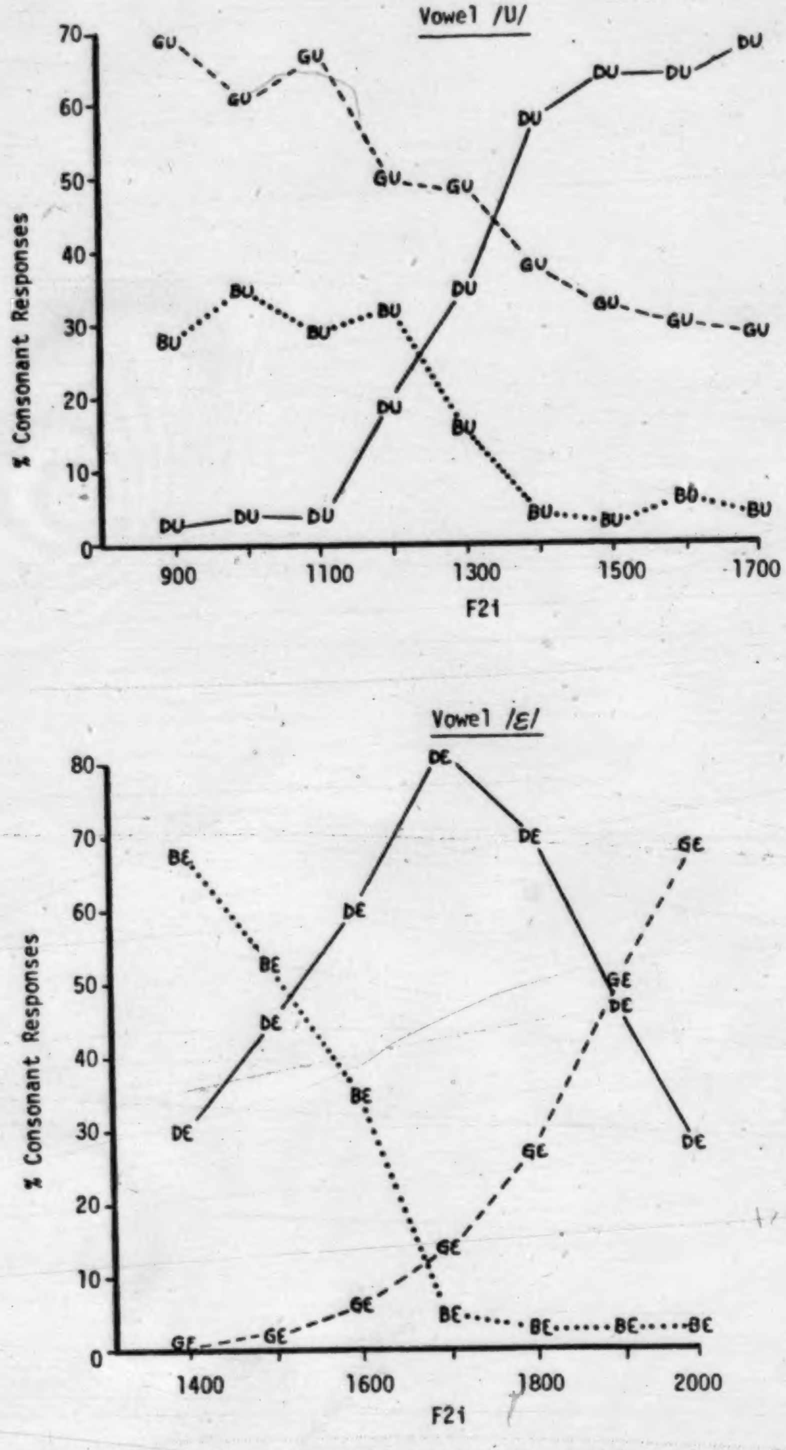


FIGURE 4.15 F2i Marginals; Non-ambiguous Vowels /U/ and /ɛ/; 8 Subjects x 5 Repetitions

perception, it should have manifested itself for any subject group, with or without replications.

The next section describes a subsequent test which was done with five subjects, analyzed separately, and more replications per subject. Of particular interest was the specific control of vowel response, since, in the experiments reported thus far, vowel ambiguity was a result of 'uncontrolled' order of presentation effects (i.e., stimuli were simply randomized).

## 4.3 Subject Differences in the Influence of Vowel Label

The following test was done in order to more fully test the influence of vowel labelling on the patterns of consonant categorization for separate subjects. A sufficient number of replications were carried out on individual subjects so that the log-linear analysis could be done on each subject's data; in this case, the log-linear assumptions are more adequately met (see Bock, 1975). In this study, vowel contrast was more stringently controlled so that vowel-labelling shifts were totally systematic. Thus, pretrials of syllables having good tokens of /U/ forced the subjects subsequent vowel choice to be /ɛ/ when syllables having ambiguous [U-ɛ] were presented. Similarly, pretrials of good tokens of /ɛ/ syllables forced syllables having ambiguous [U-ɛ] to be categorized as /U/ (see Fry, Abramson, Eimas and Liberman, 1962 and Thompson and Hollien, 1970). Though the same acoustic information was being presented in the ambiguous vowel case, would, the change in vowel label affect consonant categorization, or not? Do all subjects respond in a similar manner, or are there subject differences?

### 4.3.1 Subjects

Five subjects were chosen from the 'Pilot Vowel Test' (in Chapter 3) whose vowel boundaries for [U-ɛ] were nearly identical. Subjects were chosen who altered their vowel categorizations for ambiguous [U-ɛ] from /U/ to /ɛ/, upon hearing a previous 'good token' of /U/, and, of course, who altered their vowel categorization from /ɛ/ to /U/ for this ambiguous vowel when previously presented with a 'good' /ɛ/. Four subjects initially in the study were not included since they did not meet these requirements. All subjects were graduate linguistics students with no history of speech or hearing difficulties.

### 4.3.2 Stimiuli

Three series of stimuli were synthesized, using the Klatt (1980) Synthesizer. The 'good /U/' series had F2v=1200 Hz, F1=180-480 Hz with a 20 msec ramp, and F3v=2350 Hz. Six stimuli were generated having F2i=1500 Hz and 1700 Hz fully crossed with F3i=1800 Hz, 2100 Hz and 2400 Hz with transitions moving up to F2v within 30 msec. F1 was at 120 Hz. The 'good /ɛ/' series had F2v=1750 Hz, F1=180-480 Hz with a 20 msec ramp, F3v=2350 Hz, and F0=120 Hz. Nine stimuli were generated having F2i=1850 Hz, 2000 Hz and 2150 Hz fully crossed with F3i=2200 Hz, 2400 Hz and 2600 Hz with a 30 msec ramp.

The 'ambiguous' [U-ɛ] series consisted of 16 stimuli, with F2v=1600 Hz. (This F2 value was close to the vowel boundary for all five subjects.) F0 was set at 120 Hz, F1=180-480 Hz with a 20 msec ramp, and F3v=2350 Hz. The values of F2i were 1700 Hz, 1800 Hz, 1900 Hz and 2000 Hz and were fully crossed with four F3i values: 2100 Hz, 2200 Hz, 2300 Hz, and 2400 Hz. F2 and F3 transitions were 30 msec long.

### 4.3.3 Procedure

Subjects were first presented with one of the 'good' vowel series and instructed to press appropriately labelled switches on a switch box as being either /d/ or /g/. Ten randomizations of this pretrial were given. Immediately following the pretrial, the 'ambiguous series' was given and subjects again responded by appropriately pressing /d/ or /g/ switches. Ten randomizations of the 'ambiguous series' were also given. Some subjects first heard the 'good /U/' series, followed by the 'ambiguous series', while others started out with the 'good /ɛ/' series, followed by the 'ambiguous series'. After a brief rest period, the subjects who had already had started with the 'good /U/' series now listened to the 'good /ɛ/' series, followed by the 'ambiguous series', and the other subjects listened to the 'good /U/' series followed by the 'ambiguous series'. A total of four tests were given to each subject, yielding 40 responses per cell for the 'ambiguous series' (10 randomizations each).

All subjects reported a shift in their perception of the vowel in the 'ambiguous series', without realizing that the same acoustic information had been presented both times. Under one condition, then, they reported the vowel in the 'ambiguous series' as being /U/, and in the other condition, they reported hearing the vowel as being /ɛ/. Responses of consonant categorization were automatically scored.

### 4.3.4 Results

Results are shown in Figures 4.16-4.20 for each subject, PA, RH, KT, CL, and ML, respectively. Some subject differences are apparent. Subject MD seems to be responding more to F2i changes than the other subjects. In addition, she responded with fewer /g/ responses than the other subjects. Subject PA, on the other hand, responded with fewer /d/'s than did other subjects.

Figures 4.21-4.25 show the marginals of F2i with F3i pooled for each subject. The influence of vowel labelling can be more clearly seen in these figures. Subjects PA and RH (Figure 4.21 and Figure 4.22) show no difference in consonant categorization depending upon vowel categorization. Subjects KT and CL (Figure 4.23 and Figure 4.24) show only minor differences of consonant

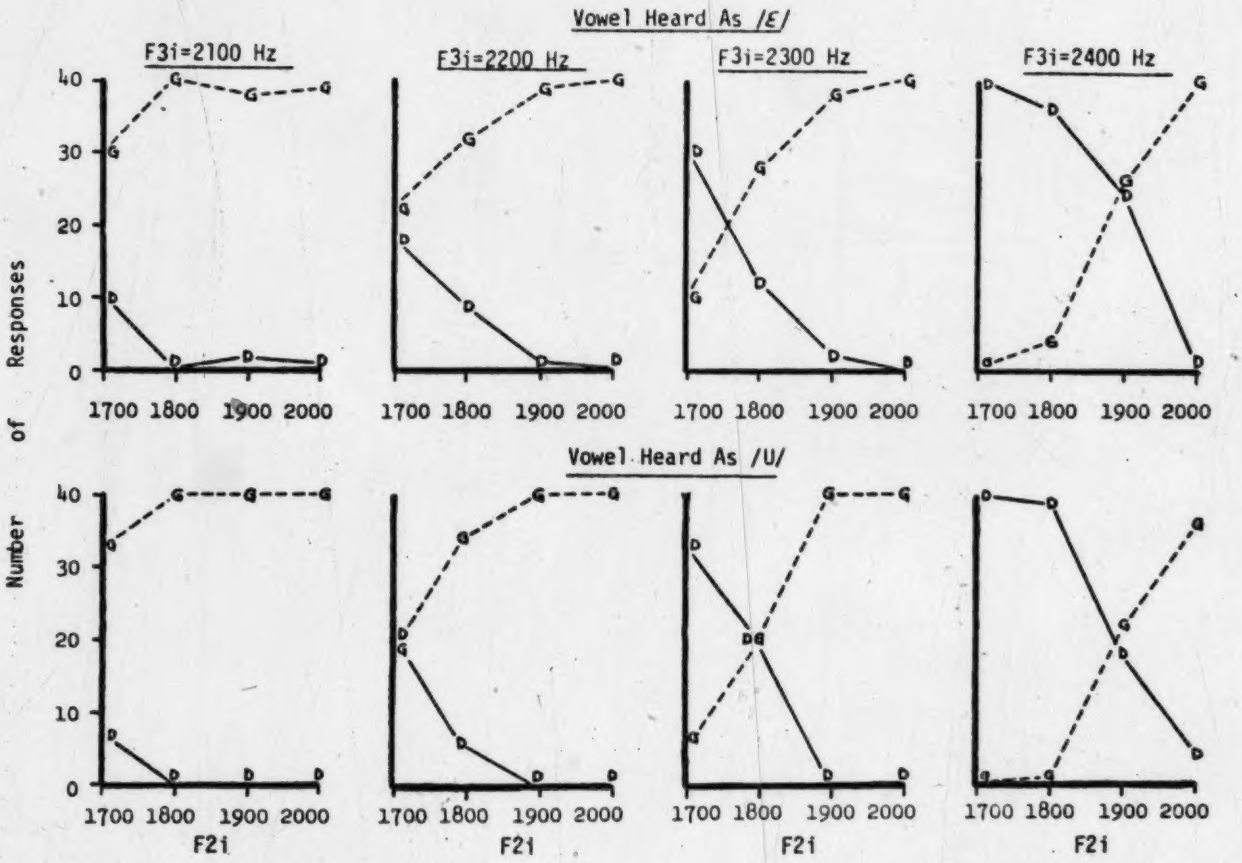


FIGURE 4.16 Consonant Categorization For Vowel Label Experiment; Subject PA

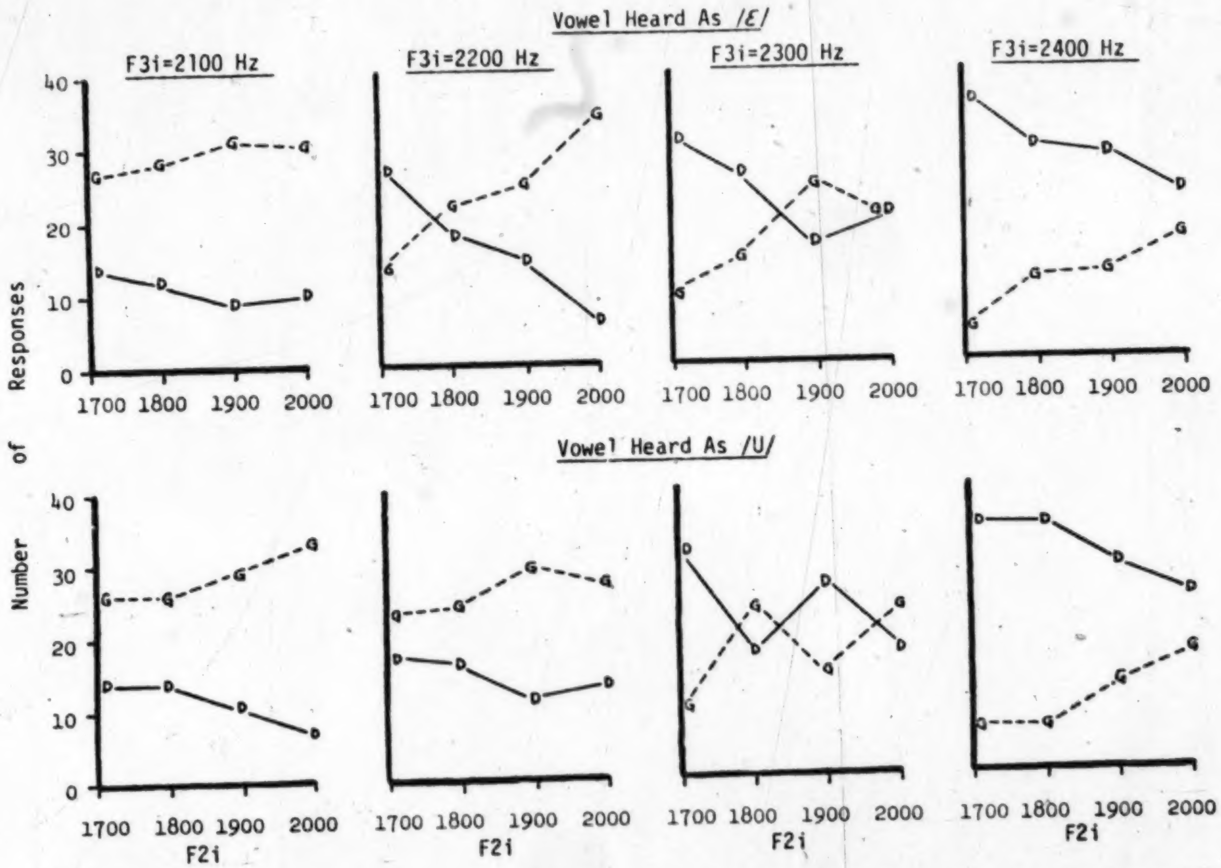


FIGURE 4.17 Consonant Categorization For Vowel Label Experiment; Subject RH

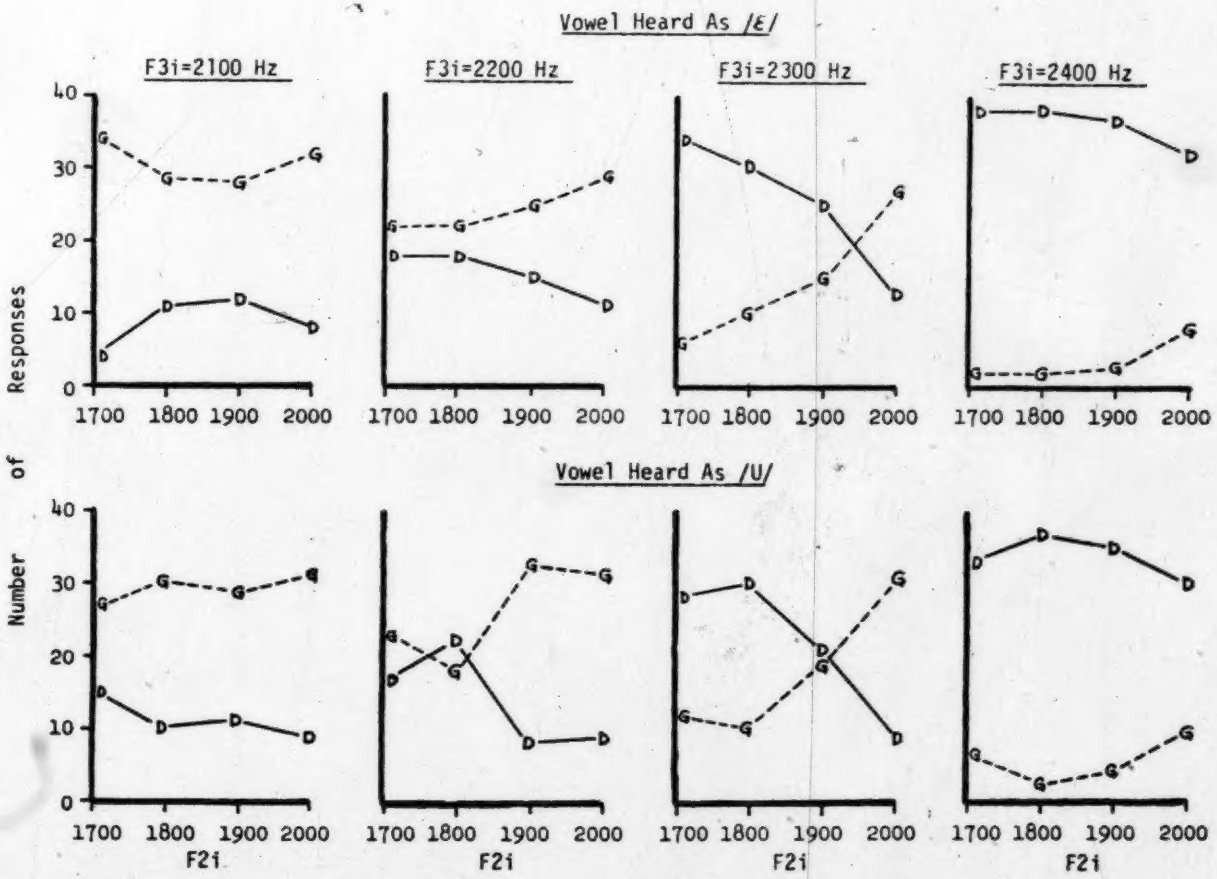


FIGURE 4.18 Consonant Categorization For Vowel Label Experiment; Subject KT

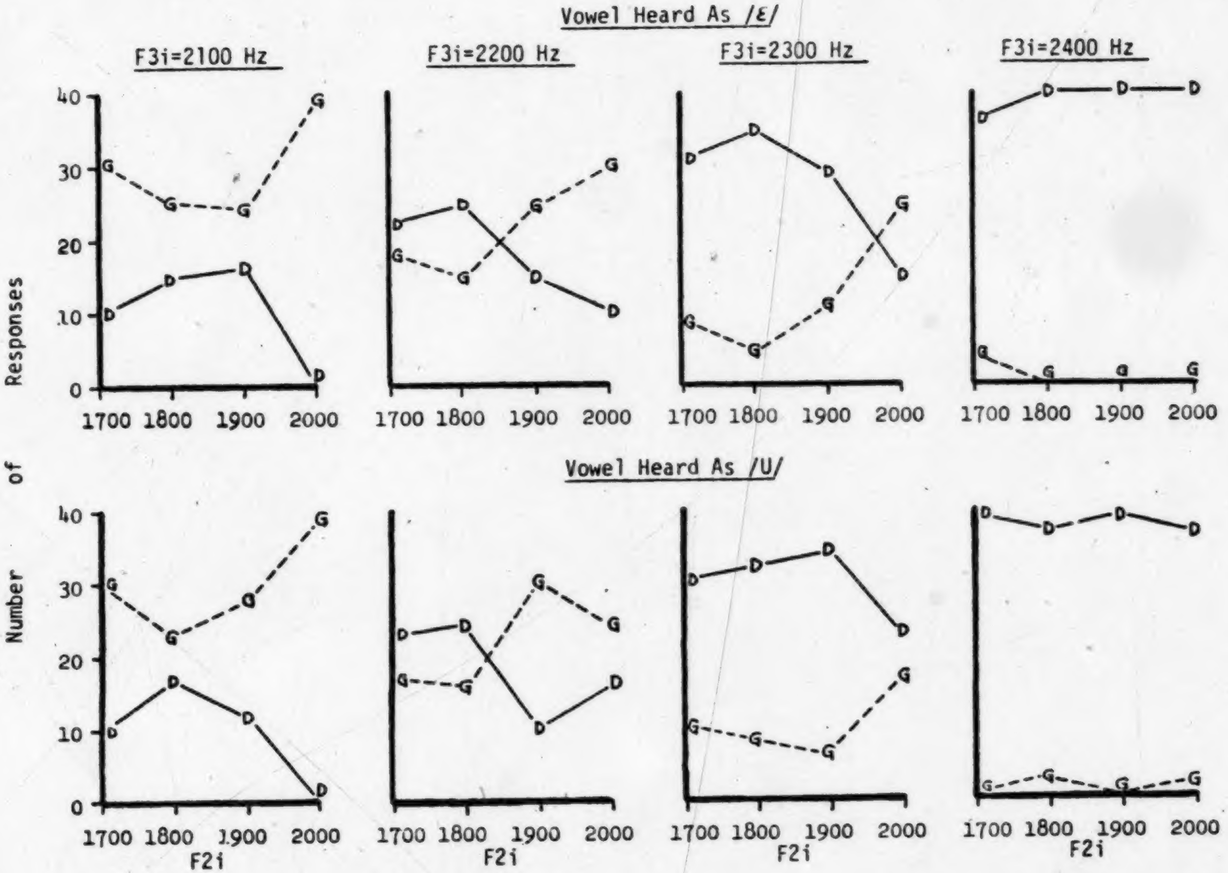


FIGURE 4.19 Consonant Categorization For Vowel Label Experiment; Subject CL

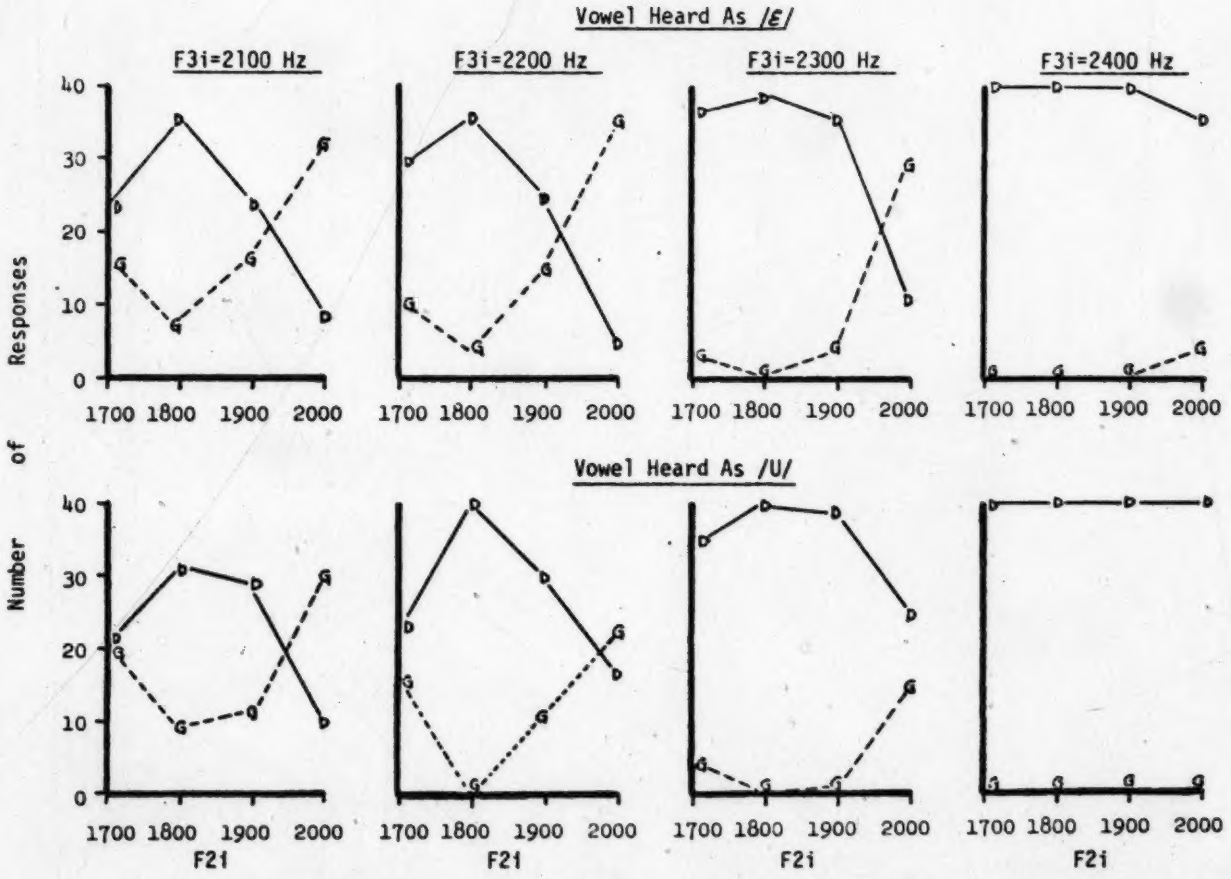


FIGURE 4.20 Consonant Categorization For Vowel Label Experiment; Subject MD

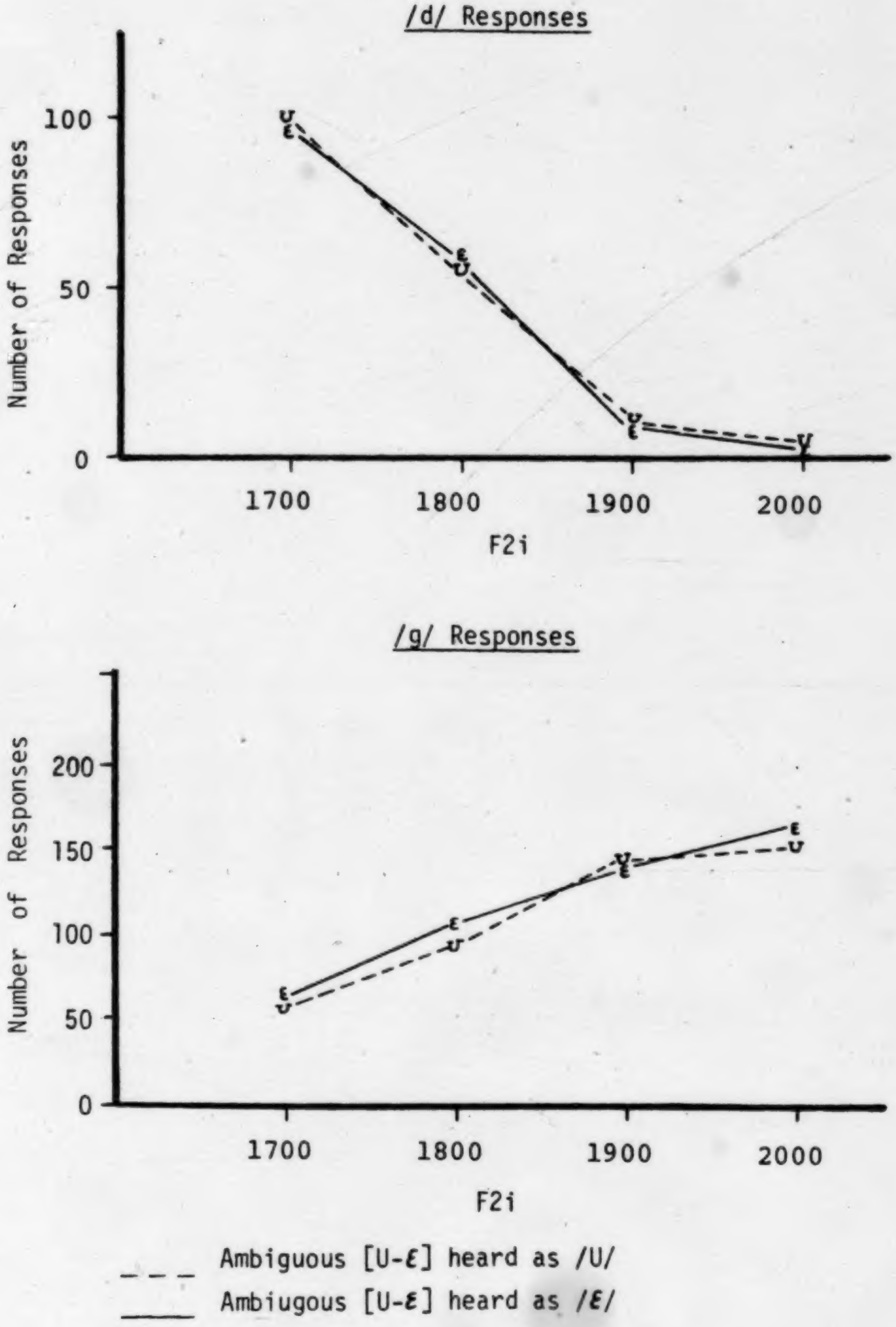


FIGURE 4.21 Marginal of F2i; Vowel Label Experiment; Subject PA

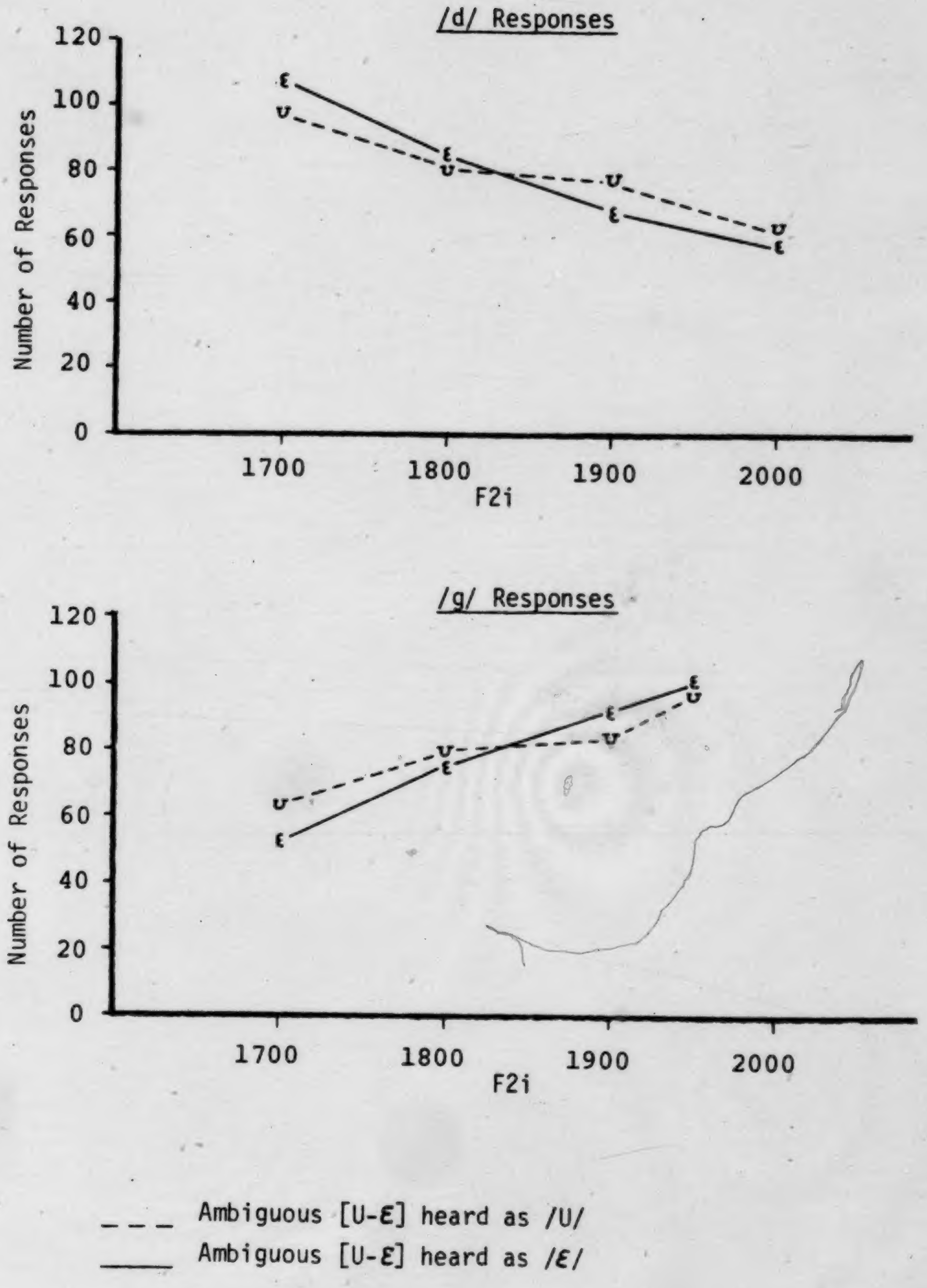


FIGURE 4.22 Marginal of F2i; Vowel Label Experiment; Subject RH

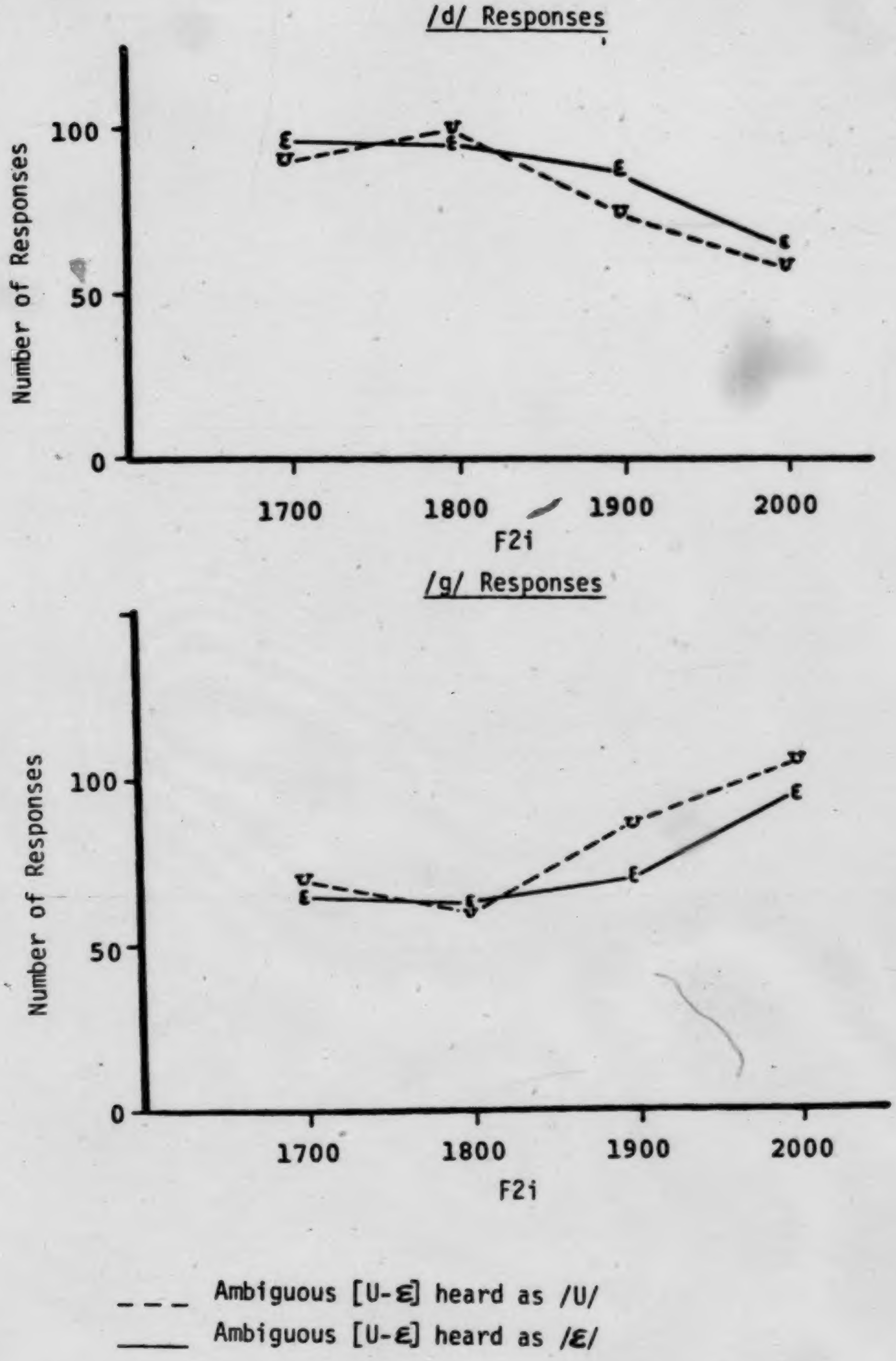


FIGURE 4.23 Marginal of F2i; Vowel label Experiment; Subject KT

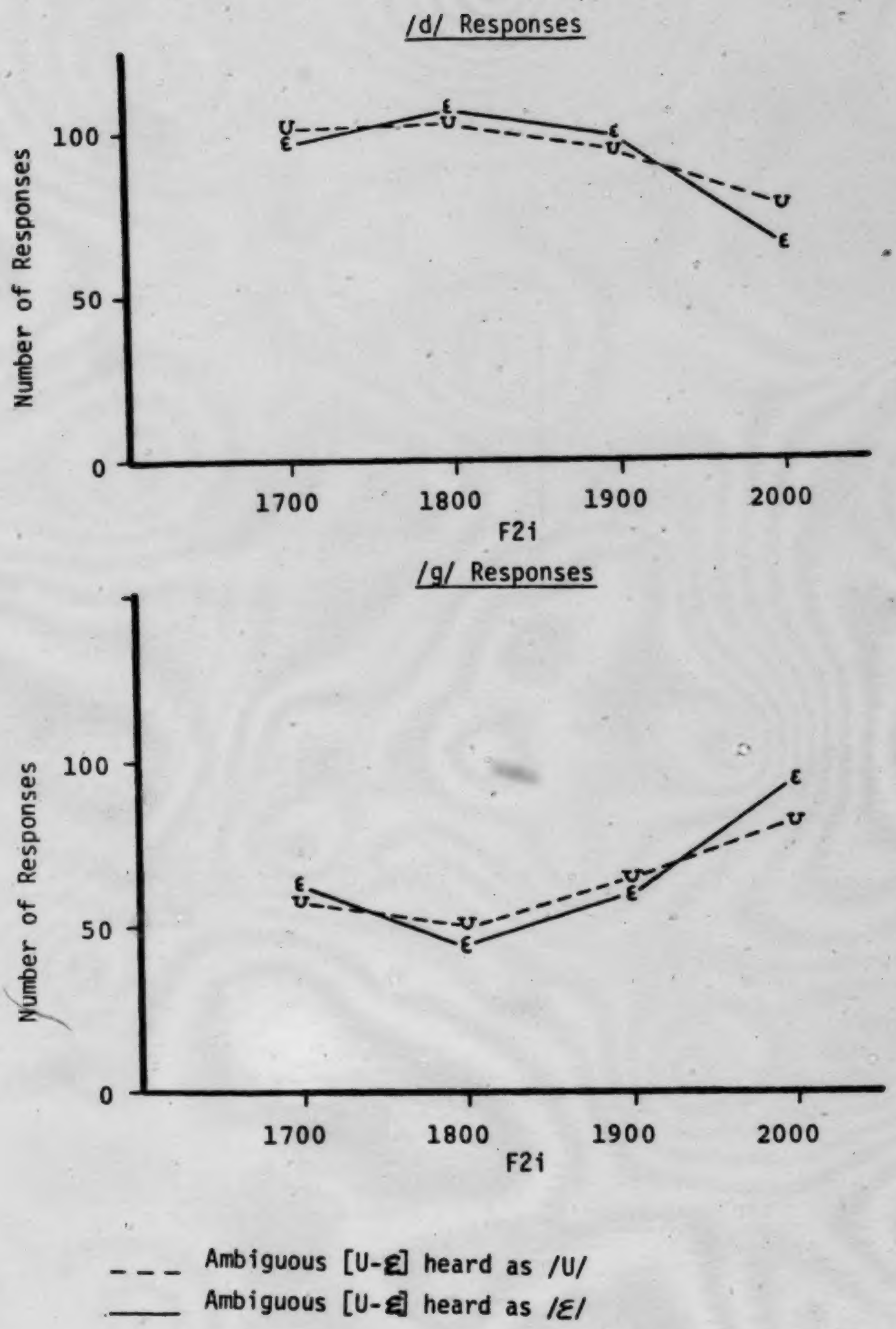


FIGURE 4.24 Marginal of F2i; Vowel Label Experiment; Subject CL

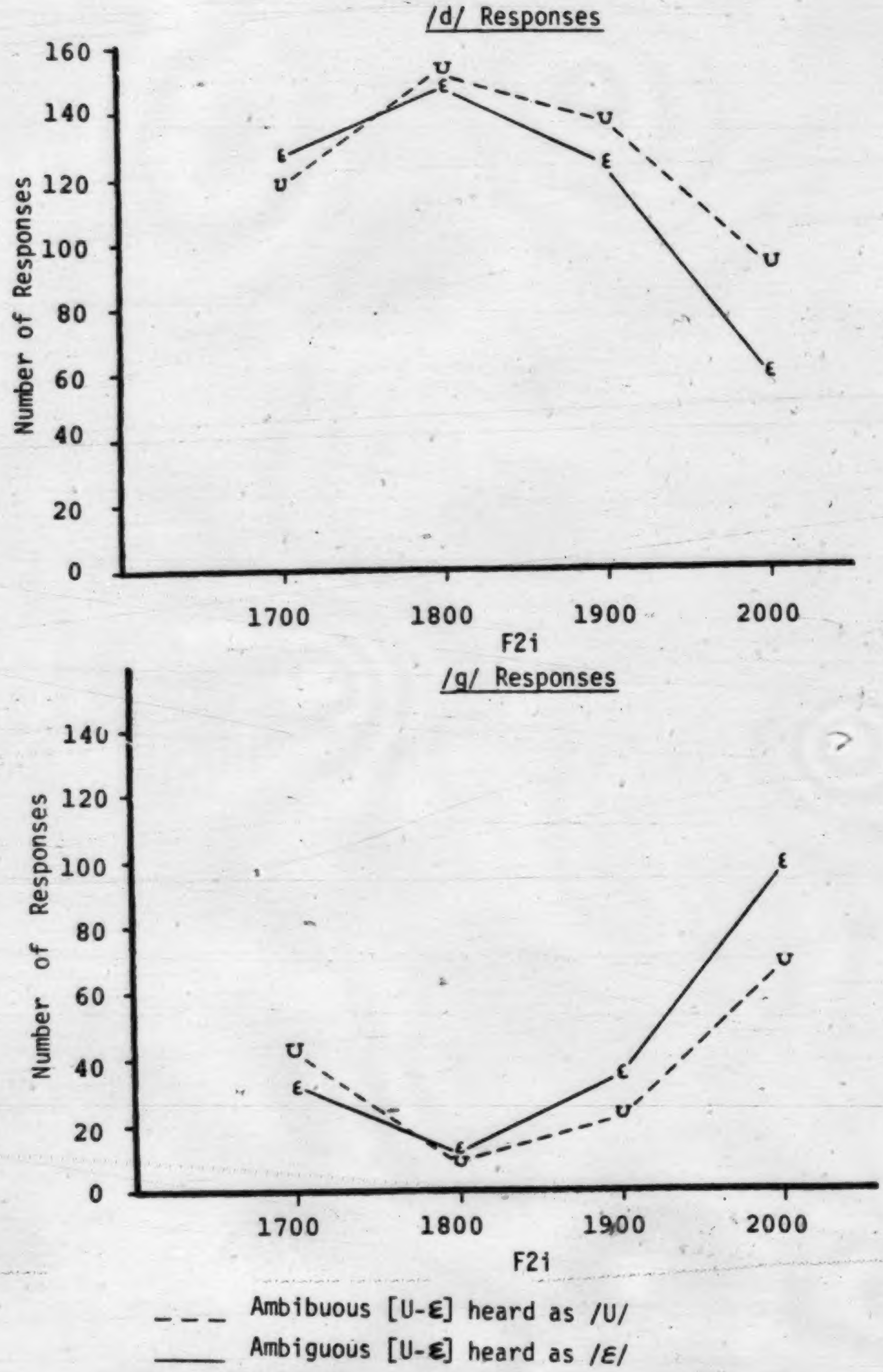


FIGURE 4.25 Marginal of F2i; Vowel Label Experiment; Subject MD

categorization when the vowel is labelled differently; KT tends to label higher F2i's as being /g/ when the vowel is heard as /U/ and as being /d/ when the vowel is heard as /ε/. CL, on the other hand, tends to the opposite direction, namely, for the highest F2i (F2i=2000 Hz), she tends to label the consonant as /d/ when the vowel is heard as /U/ and /g/ when the vowel is heard as /ε/. Subject MD (Figure 4.25) shows the most influence of vowel labelling on consonant categorization; this subject also tends to label high F2i's as /d/ when the vowel is heard as /U/ and as /g/ when the vowel is heard as /ε/.

If, in fact, vowel labelling had a strong influence in consonant categorization, we would expect that all subjects should be influenced in the same way by a change in vowel category. However, we see that subjects behave differently. Some subjects are not influenced by the vowel label at all, while other subjects seem to be at least somewhat influenced by the label. Among those subjects who do show some slight shift in consonant categorization due to vowel category changes, the trends of categorization are different between subjects. Some subjects tend to label higher F2i's as /g/ when the following vowel is heard as /ε/, but as /d/ when the following vowel is heard as /U/, while others do the reverse (i.e., label higher F2i's as /d/ when the following vowel is heard as /ε/, but as /g/ when the following vowel is heard as /U/).

Log-linear analyses were done separately for each subject's data in order to see if these consonant categorization shifts were statistically significant. The assumptions of the log-linear analysis are better met in this case since there were no 'repeated measures' in the design. Models were fit to the data which specifically tested the addition of a vowel labelling factor on consonant response patterns: either an F2i x Cononant x Vowel Label interaction term or an F3i x Consonant x Vowel Label interaction term was added to several models to see if the goodness of fit of the 'original' model significantly improved with the addition of this factor. Recall that these interactions test specifically for the change in response patterns when the 'ambiguous' vowel is labeled as one vowel compared to the situation when the 'ambiguous' vowel is labeled as another vowel. The results are shown in Table 4.5.

The log-linear analysis indicates that only one subject, MD, shows statistically significant changes in her consonant categorization when the vowel is labelled differently. Thus, though the same acoustic stimuli had been presented, this subject apparently altered her responses according to which label she had attached to the following vowel. All other subjects' consonant categorizations did not significantly change, regardless of how they labelled the following vowel. Thus, four out of the five subjects seem to be operating at an acoustic level of perception, while one

## TABLE 4.5

## Log-Linear Analysis Of Vowel-Label Experiment

Models Fit to the Data:

1. FG, FL, CL, FC
2. FG, FL, CL, FC + FCL
3. FG, GL, CL, GC
4. FG, GL, CL, GC + GCL
5. FGL, GC, FC
6. FGL, GC, FC + CL
7. FGL, GC, FC, CL + FCL
8. FGL, GC, FC, CL + GCL
9. FGL, FGC, FL, GL
10. FGL, FGC, FL, GL + CL
11. FGL, FGC, FL, GL, CL + FCL
12. FGL, FGC, FL, GL, CL + GCL

Where F = F2i, G = F3i, C = Consonant /d/ or /g/,
and L = Vowel Label /U/ or /Ɛ/.

| | | Subject | | | | |
|---|---|---|---|---|---|---|
| Model | D.F. | CL | KT | PA | RH | MD |
| | | $G^2$ | $G^2$ | $G^2$ | $G^2$ | $G^2$ |
| 1 VS. 2 | 3 | 2.56 | 1.61 | 2.06 | 2.89 | 12.39* |
| 3 VS. 4 | 3 | 2.07 | 4.18 | 3.93 | 1.26 | 7.33 |
| 5 VS. 6 | 1 | 0.12 | 2.45 | 1.48 | 0.13 | 8.78* |
| 6 VS. 7 | 3 | 4.04 | 2.19 | 2.44 | 3.30 | 19.71* |
| 6 VS. 8 | 3 | 2.17 | 4.36 | 7.06 | 1.33 | 7.74 |
| 9 VS. 10 | 1 | 0.13 | 2.52 | 1.52 | 0.14 | 8.72* |
| 10 VS. 11 | 3 | 4.52 | 2.15 | 2.50 | 3.45 | 18.30 |
| 10 VS. 12 | 3 | 2.21 | 4.63 | 7.53 | 1.34 | 8.81 |

* indicates significance to the .05 level

subject seems to be influenced by the label associated with acoustic information.

If, for subject MD, the vowel label affected consonant categorization, was this change in consonant response a type of 'heirarchical syllable-coding'? Referring back to Figure 4.4 (the syllabic templates), we can see that the onset of F2i=1800 Hz would be closer to the onset of the /dε/ template than the onset of the /gε/ template if the vowel was first identified as /ε/, and would be closer to the onset of the /dU/ template than to the onset of the /gU/ template if the vowel was first identified as /U/. In fact, there were more /g/ responses when the vowel was labelled /ε/ and more /d/ responses when the vowel was labelled /U/. The templates predict that there should have been more /d/ responses when the vowel was labelled /ε/, and not more /g/ responses, as was the case; the templates also predict that there should have been more /d/ responses when the vowel was labelled /U/, which was obtained. Thus, template predictions are correct in one case, but incorrect in the other case. Thus, this subjects responses cannot be adequately explained in terms of a hierarchical syllable based unit.

In fact, the stimulus value that could be a 'test case' of hierarchical phonetic decisions did not affect *any* subjects response categorizations in the direction predicted by the onsets of the syllabic templates of Figure 4.4; the onset value of F2i=1600 Hz is closer to the onset of template /dε/ than to the onset of template /gε/ if the vowel is first identified as /ε/ but closer to the onset of the template /gU/ than to the onset of the template /dU/ if the vowel is first identified as /U/.

Mermelstein (1978) also reported some subject differences in his experiment on the effect of vowel labelling on _final_ consonant categorization, when using duration as the variable stimulus parameter. However, he tested for a category by category interaction, which could also be interpretted as a response bias towards a particular syllable. In his study, eight out of ten subjects showed no significant vowel labelling effect; the two subjects who did show significant vowel and consonant label interactions had opposite trends of response. It is interesting to note that in Mermelstein's study, though some subjects showed a vowel label effect, *most* subjects did not. The study reported here also showed that most subjects categorization response patterns were not affected by vowel label.

Mermelstein also demonstrated that vowel label effects could not be accounted for on the basis of syllabic coding, since the subjects who showed vowel label effects gave opposite trends of responses. Similarly, this study demonstrated that the vowel label effects on consonant categorization response patterns are not easily accountable in terms of 'syllabic-coding'.

Massaro and Cohen (1983) also showed that a model which had a three-way interaction of Category x Category x Stimulus Characteristic in the perception of initial consonant clusters was not an improvement over a model which only had a two-way Category x Category interaction. Thus, this study corroborates Massaro and Cohen's findings.

In summary, this study showed that vowel label decisions did not influence the majority of subjects consonant category response patterns. The responses of the one subject that did show some vowel label influence in consonant categorization reponse patterns could not be explained in terms of hierarchical syllabic unit processing. Since there were only five subjects involved in the study, it is impossible to generalize to the population as a whole. However, this study has pointed out that there may be different listening strategies, and that pooling all subjects' data can hide some interesting aspects of perceptual behavior.

## 4.4 Summary

Pooled subject data indicated that subjects' consonant response patterns were affected by the phonetic labelling of the following vowel for the ambiguous [U-ɛ] vowel set. When the ambiguous vowel was labelled /ɛ/, response patterns looked similar to those obtained for the /ɛ/ vowel set; when the ambiguous vowel was labelled /U/, however, response patterns were not similar for /b/ responses, but were for /d/ and /g/ responses.

Results of data from eight subjects with five replications showed no significant interaction effects between consonant response, vowel response and a stimulus characteristic.

When the vowel category response is more stringently controlled, most subjects show no vowel label effect on their response patterns, and thus seem to be attending to the acoustic parameters of the signal, regardless of phonetic label. The responses of the one subject who did show an effect of vowel label on her consonant responses could not be accounted for on the basis of onsets of formant-based syllable templates. This indicates that the acoustic parameters discussed in Chapter 3 can be assessed without a priori categorization of the vowel context.

Further research is required, however, particularly across language systems. Cross-linguistic tests could provide interesting insights into the question of 'phonetic' vs. 'auditory' perceptual processes. It would be interesting to test vowel label effects in five-vowel language systems so that the F2v ranges over only two vowel categories (a back vowel and a front vowel), rather than three vowel categories as is the case in English.

## 5. FORMANT AND SPECTRAL SHAPE CUES IN NATURAL SPEECH

In the preceeding chapters, the role of formant and spectral shape information were compared for synthetic stimuli. In this chapter, natural speech is analyzed in order to see if the features that appeared to be important in the perception of the synthetic stimuli are evident in natural speech.

The raw experimental data was collected by linguistics students enrolled in a graduate phonetics course at the University of Alberta. This data is reanalyzed in terms of formant and spectral information. The first section will describe the role of onset spectral characteristics when brief portions of the stimulus onset was presented; the second section will describe the roles of formant and spectral information when response shifts were evident as more of the signal was appended.

### 5.1 Onset Spectral Characteristics

#### 5.1.1 Subjects

Five native English listeners with linguistic training were used in this experiment. None of the subjects had a history of hearing difficulties.

#### 5.1.2 Stimuli

Spoken tokens from Chapter 2 of four speakers (two male, PFA and RAH, and two female, TMD and MLD) were windowed so that 4 msec of the signal after the initial burst had full amplitude; an additional 8 msec of the signal was included which was windowed by a linear offramp. Thus, 12 msec of the signal was played to listeners, but only the first 4 msec had full amplitude. Stimuli were presented in a quiet laboratory.

### 5.1.3 Procedure

Subjects heard two randomizations of these stimuli in separate sessions. They were instructed to press appropriately marked switches connected to the PDP-12 if they heard a /p/ or /b/, /t/ or /d/, and /k/ or /g/. Subjects were allowed to hear the stimuli as often as they wished and thus controlled the rate of presentation.

### 5.1.4 Results

Table 5.1 shows the correct identification rates of /b/, /d/ and /g/ for each vowel, averaged over speaker, listener and trials. Results show that /b/ is more poorly recognized than either /d/ or /g/ for these brief stimuli. /g/'s, however, are well identified. This is rather surprising in light of the fact that Steven's and Blumstein found that, for brief synthetic stimuli, /b/'s were well-identified, while /g/'s were poorly-identified.

Table 5.1 also shows the effect of vowel context on identification of these brief stimuli; clearly, vowel context is influencing the identification rates. /b/'s

## TABLE 5.1

## Correct Identification of Natural Data

Stimulus Length

| | 4 msec | | | 8 msec | | | 16 msec | | | 32 msec | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | b | d | g | b | d | g | b | d | g | b | d | g | $\bar{x}$ |
| i | 74.6 | 98.2 | 75.3 | 84.6 | 97.9 | 82.3 | 97.9 | 98.7 | 87.8 | 97.5 | 100 | 98.1 | 91.1 |
| I | 68.6 | 90.4 | 85.5 | 93.7 | 91.8 | 94.0 | 93.5 | 98.0 | 96.9 | 85.3 | 100 | 99.1 | 91.4 |
| e | 63.7 | 96.5 | 85.5 | 87.6 | 97.7 | 94.3 | 95.1 | 99.0 | 99.1 | 91.7 | 99.6 | 100 | 92.5 |
| ɛ | 63.6 | 98.0 | 88.5 | 76.3 | 98.4 | 96.9 | 75.0 | 100 | 98.0 | 79.0 | 100 | 98.1 | 89.3 |
| æ | 89.0 | 76.7 | 96.9 | 91.0 | 84.2 | 97.3 | 92.8 | 95.7 | 96.9 | 94.5 | 98.7 | 97.8 | 92.6 |
| ə | 93.3 | 95.7 | 95.8 | 96.0 | 98.7 | 99.6 | 99.1 | 98.7 | 99.6 | 99.6 | 100 | 98.7 | 97.9 |
| ɔ | 95.8 | 96.9 | 98.4 | 96.0 | 99.3 | 99.6 | 99.6 | 98.6 | 100 | 99.1 | 100 | 99.0 | 98.5 |
| o | 91.2 | 97.0 | 98.2 | 98.8 | 98.7 | 99.3 | 100 | 99.3 | 99.6 | 99.6 | 99.1 | 96.8 | 98.1 |
| U | 88.9 | 84.9 | 99.1 | 96.2 | 88.3 | 100 | 100 | 100 | 99.6 | 99.6 | 100 | 99.6 | 96.3 |
| u | 88.1 | 91.9 | 97.5 | 89.3 | 99.6 | 99.0 | 96.8 | 99.6 | 100 | 93.5 | 100 | 97.2 | 96.0 |
| ɚ | 95.0 | 87.2 | 94.3 | 97.2 | 91.7 | 98.8 | 99.3 | 100 | 99.6 | 99.1 | 99.6 | 100 | 96.8 |
| $\bar{x}$ | 82.9 | 92.1 | 92.3 | 91.5 | 95.1 | 96.5 | 95.4 | 98.9 | 97.9 | 94.4 | 99.7 | 98.6 | |
| $\bar{D}$ | $\bar{D}_1$=89.1 | | | $\bar{D}_2$=94.4 | | | $\bar{D}_3$=97.4 | | | $\bar{D}_4$=97.6 | | | |

before front vowels are not well-identified (72% correct average identification), but /b/'s before back and central vowels are well-identified (92% correct average identification). /g/'s before high and mid-high vowels are identified better than /g/'s before back vowels.

The low error rates of these brief /g/ stimuli indicate that Stevens and Blumstein's idea that a longer time (i.e. more than 40 msec) is needed to 'build up' a representation of spectral compactness may not be necessary; similarly, Kewley-Port's feature of 'compact through time' seems to be unnecessary.

### 5.1.5 Analysis

The onset spectra are shown in Appendix 1. An analysis of spectral shape properties was done using Stevens and Blumstein onset spectral shape templates (1979), and features from Kewley-Port (1980) which pertain to the onset spectral shape.

#### 5.1.5.1 Stevens and Blumstein Templates

Stevens and Blumstein templates were fit to the onsets of each of these brief stimuli. Results are shown in Tables 5.2-5.5 for each of the speakers, PFA, RAH (two males) TMD and MLD (two females), respectively. Pluses indicate that the spectra fit the template, minuses indicate that the spectra did not fit the template.

## TABLE 5.2

### Natural Data (4 msecs onset); Stevens & Blumstein Templates;

### Subject PFA

Front Vowel Context

|  |  | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | compact | + | - | - | - | - |
|  | diffuse rising | - | + | + | + | - |
|  | diffuse falling | - | - | - | - | + |
|  | dominant response | /b/ | [b-d] | [b-d] | [b-d] | /b/ |
| D | compact | - | - | - | - | - |
|  | diffuse rising | + | + | + | + | + |
|  | diffuse falling | - | - | - | - | - |
|  | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | compact | - | + | + | + | + |
|  | diffuse rising | - | - | - | - | - |
|  | diffuse falling | - | - | - | - | - |
|  | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

Central Or Back Vowel Context

|  |  | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | compact | - | - | - | - | - | - |
|  | diffuse rising | + | - | + | + | + | - |
|  | diffuse falling | - | + | - | + | - | + |
|  | dominant response | /b/ | /b/ | [b-d] | [b-d] | /b/ | /b/ |
| D | compact | - | - | + | + | + | + |
|  | diffuse rising | + | + | + | - | - | - |
|  | diffuse falling | - | - | - | - | - | - |
|  | dominant response | /d/ | /d/ | /d/ | [d-g] | [d-g] | [d-g] |
| G | compact | + | - | ++ | + | - | ++ |
|  | diffuse rising | - | - | - | - | - | - |
|  | diffuse falling | + | + | + | - | + | - |
|  | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate spectral shapes are: diffuse falling for /b/, diffuse rising for /d/, compact for /g/

## TABLE 5.3

Natural Data (4 msecs onset); Stevens & Blumstein Templates;

Subject RAH

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | compact | - | - | - | - | + |
| | diffuse rising | + | + | + | + | - |
| | diffuse falling | - | - | - | + | + |
| | dominant response | [b-d-g] | [b-d-g] | [b-d] | [b-d-g] | [b-d] |
| D | compact | + | - | + | + | + |
| | diffuse rising | + | + | + | + | + |
| | diffuse falling | - | - | - | + | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | compact | + | + | + | + | ++ |
| | diffuse rising | - | - | - | - | - |
| | diffuse falling | - | - | - | - | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | compact | - | - | - | - | + | - |
| | diffuse rising | + | - | + | - | - | - |
| | diffuse falling | - | + | - | + | + | + |
| | dominant response | /b/ | /b/ | /b/ | /b/ | /b/ | /b/ |
| D | compact | - | - | - | + | - | + |
| | diffuse rising | + | + | + | + | + | + |
| | diffuse falling | - | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | compact | - | ++ | + | ++ | ++ | + |
| | diffuse rising | - | - | - | - | - | - |
| | diffuse falling | + | - | - | - | - | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | [d-g] |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate spectral shapes are: diffuse falling for /b/, diffuse rising for /d/, compact for /g/

## TABLE 5.4

Natural Data (4 msecs onset); Stevens & Blumstein Templates;

Subject TMD

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | compact | + | - | + | - | - |
| | diffuse rising | - | + | + | + | - |
| | diffuse falling | + | + | + | - | - |
| | dominant response | [b-d-g] | [b-d] | [b-d] | [b-d] | /b/ |
| D | compact | + | + | + | - | + |
| | diffuse rising | + | + | - | + | - |
| | diffuse falling | - | - | - | + | - |
| | dominant response | /d/ | [d-g] | /d/ | /d/ | [d-g] |
| G | compact | + | + | + | + | ++ |
| | diffuse rising | - | - | - | - | - |
| | diffuse falling | - | - | - | - | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | compact | - | - | - | - | - | - |
| | diffuse rising | + | - | - | + | - | - |
| | diffuse falling | - | + | + | - | + | + |
| | dominant response | /b/ | /b/ | /b/ | /b/ | /b/ | /b/ |
| D | compact | + | - | - | + | + | - |
| | diffuse rising | + | + | + | + | + | + |
| | diffuse falling | + | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | compact | + | ++ | + | + | ++ | + |
| | diffuse rising | - | - | - | - | - | + |
| | diffuse falling | - | + | - | - | - | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate spectral shapes are: diffuse falling for /b/, diffuse rising for /d/, compact for /g/

5

## TABLE 5.5

### Natural Data (4 msecs onset); Stevens & Blumstein Templates;

### Subject MLD

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | compact | - | - | + | - | + |
| | diffuse rising | + | + | + | + | - |
| | diffuse falling | + | - | - | - | - |
| | dominant response | /b/ | /b/ | [b-d] | [b-d] | /b/ |
| D | compact | + | + | - | + | - |
| | diffuse rising | + | + | + | + | + |
| | diffuse falling | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | [d-g] |
| G | compact | + | + | + | + | + |
| | diffuse rising | - | - | - | - | - |
| | diffuse falling | - | - | - | - | - |
| | dominant response | [d-g] | [d-g] | [d-g] | [d-g] | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | compact | + | - | - | - | - | - |
| | diffuse rising | + | + | - | + | + | + |
| | diffuse falling | - | - | + | - | - | - |
| | dominant response | [b-d] | /b/ | /b/ | /b/ | [b-d-g] | /b/ |
| D | compact | - | - | + | - | - | - |
| | diffuse rising | + | + | + | + | + | + |
| | diffuse falling | - | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | compact | ++ | ++ | ++ | + | + | - |
| | diffuse rising | - | - | - | - | - | - |
| | diffuse falling | - | - | - | - | - | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate spectral shapes are: diffuse falling for /b/, diffuse rising for /d/, compact for /g/

As can be seen in Tables 5.2-5.5, *two* of the Stevens and Blumstein templates often fit the onset spectra of these stimuli (e.g., compact and diffuse rising templates both fit the onset spectral shape of /d/ in the context /o/); thus, although there is correct acceptance of a particular template, there is also a high incidence of false alarms.

Some of the onset spectra fit both the diffuse rising and diffuse falling templates. Clearly, the notion of 'diffuse rising' and 'diffuse falling' should be redefined. It seems inappropriate to have two seemingly conflicting spectral templates *both* fit these spectra.

A closer examination of the /d/ and /b/ spectra in Appendix I reveals that many of the spectra actually both rise and fall in a type of 'hump shape'. However, the peak points in the spectra are at differing frequency locations for /d/ and /b/ spectra and, furthermore, the slopes of the rise and fall differ for these two stop categories. /b/ spectra seem to have a gradual sloping rise to a peak, and then have a fairly steep drop-off of high frequency energy. (This is consistent with the feature 'grave', i.e., presence of low frequency energy, proposed by Jakobson, Fant and Halle, 1952.) /d/'s, on the other hand, have a sharp rise to a frequency location often situated at or near the 'locus' value (Delattre, et al., 1955), with only a small sloping drop-off in high frequency energy. (This is consistent with the feature 'acute', i.e., presence of high frequency energy, proposed by Jakobson et al., 1952.)

In addition, the compact spectrum for /g/'s (especially for /b/'s before back vowels) often display *two* local 'prominent peaks', one in the mid-frequency range described by Kewley-Port (1980) and one at approximately 5000-6000 Hz, so that two of the compact templates fit in two local areas of the spectrum.

One additional aspect of the Steven's and Blumstein templates which need further revision is the fact that the diffuse rising template allows for a frequency peak in the locus area, but this peak often fits the compact spectral template as well. Thus the compact template falsely accepts /d/ onset spectra. Perhaps a hierarchical application of the templates could help, whereby the diffuse rising templates are first fit to the spectra, and then the compact templates are fit only to those spectra which did not fit the diffuse rising template at the extended locus area (as defined by Stevens and Blumstein, 1979).

The onset spectral shape of these stimuli can not always be related to the dominant responses of listeners; for example, some stimuli have diffuse rising shapes, but are heard as /b/'s. However, most tokens with majority /g/ responses fit the compact templates.

Thus the compact feature seems to be an important cue for velars in natural speech as well as in the synthetic speech perception experiments that were done in Chapter 3.

Results also indicate that vowel context is important. The results for each speaker are presented below.

Speaker PFA (Table 5.2)

For this speaker, onset spectral shapes for each stop category was affected by vowel context. The spectra of /d/ and /g/ before front vowels fit the appropriate diffuse rising and compact templates without being falsely accepted by the other templates; however, for both /d/ and /g/ in back vowel contexts, four out of six spectra were falsely accepted by other templates. /b/'s before back vowels more often had the appropriate 'diffuse falling' onset spectral shape (three out of six) than did /b/'s before front vowels (one out of five).

It can be seen that the 'fit to appropriate spectral shape templates' does not seem to fully account for the 'dominant response' of listeners to *all* of these brief stimuli. Stimuli having majority /b/ responses often do not have the 'appropriate' diffuse falling shape. Stimuli having majority /d/ responses, on the other hand, do have the appropriate diffuse rising shape. Many of the stimuli which obtained dominant /g/

responses appropriately fit the compact spectral template, though many of these stimuli were falsely accepted by other templates. Moreover, most of the stimuli which were 'ambiguous' for listeners could not be interpretted in terms of fitting two or more templates; only one ambiguous stimulus, the /bU/ token (which was ambiguously heard as [b-d]) fit both the diffuse rising and the diffuse falling templates.

A compact template often fits two local areas of the spectra for /g/'s, particularly /g/'s before back vowels.

### *Speaker RAH (Table 5.3)*

For speaker RAH (Table 5.3), the influence of vowel context is also evident. For both /b/'s and /d/'s in back vowel contexts, four out of six have the appropriate diffuse shape (i.e., diffuse rising for /d/'s and diffuse falling for /b/'s) with no false acceptances by the other templates. However, for /b/'s in front vowel context, there are no spectra which *only* fit the appropriate diffuse falling template; for /d/'s in front vowel context, only one out of the five spectra fit only the diffuse rising template. The spectral shape of /d/'s before front vowels often fit both the diffuse and compact templates, owing to prominant peaks near the 'locus' value (four out of five spectra fit both templates). The onset spectra of some /g/'s before back vowels are falsely accepted by the diffuse falling template (two out of six spectra), but there are no such 'false alarms' with /g/'s before front vowels. As with speaker PFA, the compact templates fit some of the /g/ spectra (four out of eleven) in two local areas of the spectrum, particularly for /g/'s before back vowels.

Again, the onset spectral shape characteristics seem to be unable to account for *all* of the dominant responses of listeners when presented with these brief tokens. Some of the tokens which obtained majority /b/ responses had the appropriate diffuse falling shape, while others fit the diffuse rising template. Every token which obtained a dominant /d/ response fit the diffuse rising template, but there were also many false alarms. Most tokens with majority /g/ responses fit the compact templates.

None of the ambiguous tokens fit two or more of the appropriate templates. For example, the token /gɚ/ was ambiguously heard as [d-g], but the onset spectral shape of this stimulus did not fit both the compact and diffuse rising templates; rather, it fit the compact and diffuse falling templates.

### Speaker TMD (Table 5.4)

For speaker TMD, vowel context affected the onset spectral shape. /b/'s and /d/'s before central or back vowels generally have the appropriate falling and rising diffuse spectral shape, respectively, without also fitting another inappropriate template (four out of six

/b/'s and three out of six /d/'s in this vowel context). However, there were no tokens of /b/'s and /d/'s before front vowels which had the appropriate diffuse shapes without also fitting an inappropriate template. /g/'s before front vowels were never falsely accepted by other templates, whereas /g/'s in two back vowel contexts had 'false alarms'.

The compact templates sometimes fit in two local areas of the spectrum for /g/'s (three spectra out of eleven), particularly before back vowels. In addition, this subjects /d/'s also fit the compact template especially before front vowels, where the template was picking up the peak near the 'locus' value.

The Stevens and Blumstein template fits do not reflect dominant response categories of listeners for all tokens. Some stimuli which obtained majority /b/ responses had the appropriate diffuse falling onset spectra, while others had diffuse *rising* spectra. Most stimuli which obtained majority /d/ responses fit the diffuse rising template but also were falsely accepted by other templates (though, as mentioned above, some of these false alarms were due to a compact template fitting a peak near the 'locus' value). Stimuli which obtained majority /g/ responses *all* fit a compact template and, in some cases, a compact template fit two local areas of the spectrum. Only two /g/ tokens were falsely accepted by other templates.

Stimuli which were ambiguous under this short presentation condition did *not*, in general, fit two or more of the appropriate spectral shape templates. Only one ambiguous token, the syllable /bI/ (which was heard ambiguously as a /d/ or a /b/) fit both the diffuse rising and the diffuse falling templates.

### Speaker MLD (Table 5.5)

For speaker MLD (Table 5.5), onset spectral shapes were only slightly affected by vowel context. The spectra of /d/'s before front vowels had more 'false alarms' (three out of five false alarms) with the compact template than /d/'s before back vowels (one out of six false alarms); here again, the compact template was picking up the spectral peak near the 'locus' area. It is interesting to note, though, that /d/'s before front vowels often have a peak at the locus, while /d/'s before back vowels do not necessarily have this locus peak (cf. Blumstein and Stevens, 1980).

Listeners' dominant responses to these shortened tokens can not be *all* accounted for on the basis of onset spectral shape characteristics. Tokens with majority /b/ responses often fit the diffuse *rising* template, instead of the 'appropriate' diffuse falling template. Stimuli with majority /d/ responses all fit the diffuse rising template, but are often falsely accepted by the compact templates as well (particularly before front vowels). Almost all of the stimuli which obtained majority /g/ responses were accepted by a compact template, and were not falsely accepted by the other templates.

Stimuli which are ambiguously heard are also not adequately accounted for on the basis of their onset spectral shape characteristic. None of the 'ambiguous' tokens fit the appropriate templates (i.e., appropriate for both of the categories with which subjects are responding).

### Summary

The Stevens and Blumstein templates do not seem to capture the important differences in onset spectra as a function of the vowel context. More onset spectra of consonants before back vowels fit the appropriate templates without false alarms than do onset spectra of consonants before front vowels (28 of the 60 cases, or 46%, across all subjects for front vowel context vs. 39 of the 72 cases, or 53% for back and central vowel contexts). If a hierarchical application of the diffuse rising and compact templates were done, 36 of the 60 front vowel contexts, or 60%, would fit the appropriate templates without false alarms vs. 39 of the 72 back vowel contexts (64%).

In addition, these templates do not account for all of the dominant responses of listeners. In general, the compact template fits onset spectra which are identified as /g/'s by listeners, with no false alarms (30 out of

39, or 77%, of such cases). Stimuli which obtained majority /d/ responses almost always fit the diffuse rising template (36 out of 38, or 95%); however, the compact template often falsely accepts these /d/ stimuli as well (42% of the time), particularly for /d/'s before front vowels, which more often have a peak near the 'locus' value. Thus, for unambiguous /d/'s, only 20 out of 38 stimuli (or 53%) appropriately fit *only* the diffuse rising template. There were 26 stimuli which obtained dominant /b/ responses; of these, 11 had the appropriate diffuse falling shape (42%) without fitting other templates.

Some suggested revisions of the Stevens and Blumstein templates are presented below:

1. The diffuse rising template could be fit before the compact template, especially before front vowels, so that spectra having a peak at the locus /d/ value would not also be categorized as /g/'s (i.e., those spectra which fit the diffuse rising template in the locus area would not be fit with a compact template). Of those stimuli which obtained clear /b/, /d/ and /g/ dominant responses (103 cases), 57% fit the appropriate templates (with no false alarms) without hierarchical template-fitting, and 73% did so *with* hierarchical template-fitting.[1]

---

[1] Percentages of stimuli with dominant /d/ responses appropriately fitting only the diffuse rising template increased from 53% to 97% with hierarchical fitting; stimuli with dominant /g/ responses which fit only the compact

2. The diffuse rising and diffuse falling templates could be redefined in terms of peak frequency location, and spectral slope values of lines fit to the rising and falling portions of the onset spectra, so that spectra would not fit both a diffuse rising and a diffuse falling spectrum at the same time. Perhaps a 'diffuse flat' template could be devised in order to test for 'ambiguous' [b-d] stimuli.
3. The templates could be 'tuned' to context sensitive acoustic information (such as F2 transition information). This notion of 'template' indicates that some aspects of the acoustic context must be included in the templates; thus, a continuous *family* of templates (analagous to the '/g/-family' of compact templates of Blumstein and Stevens, 1979) could be devised for each stop consonant category, which would be tuned to an acoustic aspect of the vowel context, though not in a hierarchical fashion (i.e., vowel categorization information would not be required a priori).

Application of the first suggested revision yielded an improvement of the total number of stimuli which fit an appropriate template without false alarms (57% without hierarchical template-fitting, 73% with hierarchical template-fitting). Unfortunately, it is

---

'(cont'd) template increased from 77% to 79% with hierarchical fitting.

beyond the scope of the present work to implement the other two suggested revisions. Clearly, an automatic template-fitting procedure is required in order to properly assess these template revisions.

### 5.1.5.2 Kewley-Port Features

Tables 5.6-5.9 show the application of Kewley-Port features to the onsets of these short stimuli. The features of 'rising spectrum at onset', 'falling spectrum at onset' , and 'prominent mid-frequency peaks' were used in the analysis. Although Kewley-Port defined one of these features as requiring a time dimension (mid-frequency peaks extending through time), this feature was nevertheless tested in order to see if it was evident at the onset.

Results show that the Kewley-Port features do not adequately account for all of the dominant response categories. Of the 26 unambiguous /b/ cases, only two had a falling spectral onset shape (or approximately 8%). However, of the 38 unambiguous /d/ cases, 35 (or 92%) had a rising spectral shape; the other three stimuli had a rising spectral shape, but also had mid-frequency peaks. Thus, 100% of the stimuli which obtained dominant /d/ responses in fact had a rising spectral shape, but there were three false alarms. Hierarchical application of these features would raise the percentage of appropriate fits from 92% to 100%. Similarly, of the 39 unambiguous /g/ cases, none of the

## TABLE 5.6

### Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject PFA

| | | Front Vowel Context | | | | |
|---|---|---|---|---|---|---|
| | | i | I | e | ɛ | æ |
| B | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | + | + | - |
| | dominant response | /b/ | [b-d] | [b-d] | [b-d] | /b/ |
| D | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | + |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | + | + | + | + | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

| | | Central Or Back Vowel Context | | | | | |
|---|---|---|---|---|---|---|---|
| | | ʌ | ɔ | o | U | u | ɚ |
| B | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | [illegible] | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /b/ | /b/ | [b-d] | [b-d] | /b/ | /b/ |
| D | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | + |
| | dominant response | /d/ | /d/ | /d/ | [d-g] | [d-g] | [d-g] |
| G | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | + | - | - | + | - | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: $\geq$ 80% response for the category

appropriate features are: falling spectrum for /b/, rising spectrum for /d/, mid-frequency peaks for /g/

## TABLE 5.7

## Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject RAH

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - |
| | dominant response | [b-d-g] | [b-d-g] | [b-d] | [b-d-g] | [b-d] |
| D | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | [illegible] | - | - |
| | mid-frequency peaks | + | + | + | + | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | rising spectrum | - | - | + | + | - | - |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /b/ | /b/ | /b/ | /b/ | /b/ | /b/ |
| D | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | + | + | + | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | [d-g] |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate features are: falling spectrum for /b/, rising spectrum for /d/, mid-frequency peaks for /g/

## TABLE 5.8

## Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject TMD

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | rising spectrum | - | + | - | + | + |
| | falling spectrum | + | - | + | - | - |
| | mid-frequency peaks | - | - | - | - | - |
| | dominant response | [b-d-g] | [b-d] | [b-d] | [b-d] | /b/ |
| D | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | + | + | + |
| | dominant response | /d/ | [d-g] | /d/ | /d/ | [d-g] |
| G | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | + | + | + | + | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | rising spectrum | + | - | - | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /b/ | /b/ | /b/ | /b/ | /b/ | /b/ |
| D | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | + | - | + | + | - | - |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate features are: falling spectrum for /b/, rising spectrum for /d/, mid-frequency peaks for /g/

## TABLE 5.9

### Natural Data (4 msecs onset); Kewley-Port Onset Features; Subject MLD

Front Vowel Context

| | | i | I | e | ɛ | æ |
|---|---|---|---|---|---|---|
| B | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | + | - | - |
| | dominant response | /b/ | /b/ | [b-d] | [b-d] | /b/ |
| D | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | [d-g] |
| G | rising spectrum | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - |
| | mid-frequency peaks | - | + | + | + | + |
| | dominant response | [d-g] | [d-g] | [d-g] | [d-g] | /g/ |

Central Or Back Vowel Context

| | | ʌ | ɔ | o | U | u | ɚ |
|---|---|---|---|---|---|---|---|
| B | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | [b-d] | /b/ | /b/ | /b/ | [b-d-g] | /b/ |
| D | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | - | - | - | - | - | - |
| | dominant response | /d/ | /d/ | /d/ | /d/ | /d/ | /d/ |
| G | rising spectrum | + | + | + | + | + | + |
| | falling spectrum | - | - | - | - | - | - |
| | mid-frequency peaks | + | + | - | + | + | + |
| | dominant response | /g/ | /g/ | /g/ | /g/ | /g/ | /g/ |

++ indicates that the compact templates fit in two local areas of the spectrum

dominant response: ≥ 80% response for the category

appropriate features are: falling spectrum for /b/, rising spectrum for /d/, mid-frequency peaks for /g/

stimuli only had the appropriate mid-frequency peaks; however, 29 of these stimuli had both mid-frequency peaks and rising spectra (i.e., 74%). Thus, hierarchical application of the mid-frequency and rising spectrum features would raise the the percentage of appropriate fits (with no false alarms) from 0% to 74%.

In general, the Kewley-Port features were not an improvement over the Stevens and Blumstein templates; using hierarchical fitting in both cases, the Stevens and Blumstein templates appropriately fit 73% of the unambiguous stimuli whereas the Kewley-Port features fit 61% of these stimuli. In particular, the feature that was designed to separate out labials from alveolars (i.e., falling spectra) was not evident; most of these stimuli had 'rising spectra', but were categorized as /b/ by listeners. If, in fact, the rise or fall of the onset spectral shape cued the difference between /b/'s and /d/'s, listeners should have heard only two /b/'s; in fact, the stimuli with falling spectra were ambiguous.

The feature 'mid-frequency peaks' was, in fact, evident for many of the stimuli which obtained a majority /g/ response (74% of the cases). However, there were some cases which did not have a 'prominent' peak in this mid-frequency range, but rather also had a smaller peak nearby (see Appendix I); in most cases, these spectra would have fit a Stevens and Blumstein compact template.

In this section, the onset characteristics of natural speech were described. In the following section, aspects of presenting a longer portion of the signal are discussed.

### 5.2 Natural Data With Response Shifts

In this section, natural stimuli with longer portions of the signal are tested; those stimuli which obtain category response shifts are analyzed in terms of spectral and formant properties.

#### 5.2.1 Subjects and Procedure

The subjects and procedure were identical to those mentioned in the previous section.

#### 5.2.2 Stimuli

Stimuli were identical to those in the previous section, although the duration of the signal was lengthened to four duration levels. The stimuli had full amplitude values for 4, 8, 16, or 32 msec; for all stimuli, an 8 msec linear offramp was provided. Thus, the full stimuli durations were 12, 16, 32 and 40 msec

#### 5.2.3 Results and Analysis

An analysis was done only on those stimuli which obtained less than 80% correct recognition, or for those which had response shifts as more of the signal was added. A response shift is defined as a shift in response from an 'ambiguous' categorization to a majority response (i.e., the categorization of a token heard ambiguously when presented with a short stimulus vs. the categorization of this token heard unambiguously when more of the signal was appended). A token was considered to be 'ambiguously' categorized if no single category exceeded 80% correct identification (i.e., two or more categories were needed in order to obtain 80% of the consonant responses.

### 5.2.3.1 Analysis of Spectral Shape

For each of these stimuli, spectral shape information was analysed by fitting the Stevens and Blumstein templates, calculating Lahiri and Blumstein ratios, and applying Kewley-Port features. In addition, formant measurements were obtained for each of these stimuli where possible. (At these brief presentation levels, formant tracking was impossible for some stimuli due to the long initial burst portion.) Plots of these spectra are shown in Appendix 2.

The results of fitting the Stevens and Blumstein templates to these spectra are shown in Tables 5.10-5.13. The results indicate that the Stevens and Blumstein templates bear some relation to the point at which subjects are shifting their responses from one category to another as more of the signal is presented, but that they do not account for *all* response category

## TABLE 5.10

### Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject PFA

| Syllable | Stimulus Length | Dominant Response | Onset Spectral Shape | Spectral Shape at Voice Onset |
|---|---|---|---|---|
| /bI/ | 4 msecs | [b-d] | diffuse rising | diffuse rising (-VOT) |
| | 8 msecs | /b/ | diffuse rising | diffuse falling |
| /be/ | 4 msecs | [b-d] | diffuse rising | diffuse (-VOT) |
| | 8 msecs | [b-d] | diffuse rising | diffuse |
| | 16 msecs | /b/ | diffuse rising | diffuse falling |
| /bɛ/ | 4 msecs | [b-d] | diffuse rising | diffuse (-VOT) |
| | 8 msecs | [b-d] | diffuse rising | diffuse |
| | 16 msecs | [b-d] | diffuse rising | diffuse |
| | 32 msecs | [b-d] | diffuse rising | diffuse |
| /bo/ | 4 msecs | [b-d] | diffuse rising | diffuse rising (-VOT) |
| | 8 msecs | /b/ | diffuse rising | diffuse falling |
| /bU/ | 4 msecs | [b-d] | diffuse | diffuse (-VOT) |
| | 8 msecs | /b/ | diffuse | diffuse falling |
| /dU/ | 4 msecs | [d-g] | compact | compact<br>diffuse rising (-VOT) |
| | 8 msecs | [d-g] | compact | compact<br>diffuse rising (-VOT) |
| | 16 msecs | /d/ | compact | diffuse rising |
| /du/ | 4 msecs | [d-g] | compact | compact<br>diffuse rising (-VOT) |
| | 8 msecs | /d/ | compact | diffuse rising |
| /dɝ/ | 4 msecs | [d-g] | compact | compact<br>diffuse rising (-VOT) |
| | 8 msecs | [d-g] | compact | compact<br>diffuse rising (-VOT) |
| | 16 msecs | /d/ | compact | diffuse rising |

(-VOT): no voicing onset; third or fourth frame used for 'spectrum at voicing onset'

## TABLE 5.11

## Natural Data with Response Shifts;
## Stevens and Blumstein Templates;
## Subject RAH

| Syllable | Stimulus Length | Dominant Response | Onset Spectral Shape | Spectral Shape at Voice Onset |
|---|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | diffuse rising | diffuse rising * |
| | 8 msecs | /b/ | diffuse rising | diffuse rising * |
| /bI/ | 4 msecs | [b-d-g] | diffuse rising | |
| | 8 msecs | /b/ | diffuse rising | diffuse rising |
| /be/ | 4 msecs | [b-d] | diffuse rising | diffuse rising * |
| | 8 msecs | [b-d] | diffuse rising | diffuse rising * |
| | 16 msecs | [b-d] | diffuse rising | diffuse rising * |
| | 32 msecs | [b-d] | diffuse rising | diffuse rising * |
| /bɛ/ | 4 msecs | [b-d-g] | diffuse | diffuse * |
| | 8 msecs | [b-d] | diffuse rising | diffuse * |
| | 16 msecs | [b-d] | diffuse rising | diffuse * |
| | 32 msecs | [b-d] | diffuse rising | diffuse * |
| /bæ/ | 4 msecs | [b-d] | compact | compact |
| | | | diffuse falling | diffuse falling* |
| | 8 msecs | [b-d] | diffuse falling | diffuse * |
| | 16 msecs | [b-d] | diffuse falling | diffuse * |
| | 32 msecs | [b-d] | diffuse falling | diffuse * |
| /gɚ/ | 4 msecs | [d-g] | compact | compact |
| | | | diffuse falling | diffuse falling (-VOT) |
| | 8 msecs | /g/ | compact | compact |
| | | | diffuse falling | diffuse falling (-VOT) |

(-VOT) no voicing onset; third or fourth frame used for 'spectrum at voicing onset'

* voicing onset at consonant release

### TABLE 5.12

### Natural Data with Response Shifts;
### Stevens and Blumstein Templates;
### Subject TMD

| Syllable | Stimulus Length | Dominant Response | Onset Spectral Shape | Spectral Shape at Voice Onset |
|---|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | compact<br>diffuse falling | compact<br>diffuse falling |
| | 8 msecs | [b-d-g] | compact<br>diffuse falling | compact<br>diffuse falling |
| | 16 msecs | /b/ | compact<br>diffuse falling | compact<br>diffuse falling |
| /bI/ | 4 msecs | [b-d] | diffuse | diffuse |
| | 8 msecs | /b/ | diffuse | diffuse falling |
| /be/ | 4 msecs | [b-d] | compact<br>diffuse | compact<br>diffuse rising |
| | 8 msecs | /b/ | compact<br>diffuse | diffuse falling |
| /bɛ/ | 4 msecs | [b-d] | diffuse rising | diffuse rising |
| | 8 msecs | /b/ | diffuse rising | diffuse falling |
| /dI/ | 4 msecs | [d-g] | compact<br>diffuse rising | compact<br>diffuse falling (-VOT) |
| | 8 msecs | [d-g] | compact<br>diffuse rising | compact<br>diffuse falling (-VOT) |
| | 16 msecs | /d/ | compact<br>diffuse rising | diffuse rising |
| /dɝ/ | 4 msecs | [d-g] | compact | compact (-VOT) |
| | 8 msecs | [d-g] | compact | compact (-VOT) |
| | 16 msecs | /d/ | compact | diffuse falling |

(-VOT): no voicing onset; third or fourth frame used for 'spectrum at voicing onset'

## TABLE 5.13

## Natural Data With Response Shifts; Stevens and Blumstein Templates; Subject MLD

| Syllable | Stimulus Length | Dominant Response | Onset Spectral Shape | Spectral Shape at Voice Onset |
|---|---|---|---|---|
| /be/ | 4 msecs | [b-d] | compact<br>diffuse | compact *<br>diffuse |
| | 8 msecs | /b/ | compact<br>diffuse | compact *<br>diffuse |
| /bɛ/ | 4 msecs | [b-d] | diffuse rising | diffuse rising |
| | 8 msecs | [b-d] | diffuse rising | diffuse rising |
| | 16 msecs | /b/ | diffuse rising | diffuse falling |
| /bʌ/ | 4 msecs | [b-d] | diffuse | diffuse * |
| | 8 msecs | [b-d] | diffuse | diffuse * |
| | 16 msecs | /b/ | diffuse | diffuse falling * |
| /bu/ | 4 msecs | [b-d-g] | diffuse rising | compact<br>diffuse rising (-VOT) |
| | 8 msecs | [b-d-g] | diffuse rising | compact<br>diffuse rising (-VOT) |
| | 16 msecs | [b-d] | diffuse rising | compact<br>diffuse rising |
| | 32 msecs | [b-d-g] | diffuse rising | compact<br>diffuse rising |
| /dæ/ | 4 msecs | [d-g] | diffuse rising | diffuse rising (-VOT) |
| | 8 msecs | [d-g] | diffuse rising | diffuse rising |
| | 16 msecs | /d/ | diffuse rising | diffuse rising |
| /gi/ | 4 msecs | [d-g] | compact | compact (-VOT) |
| | 8 msecs | [d-g] | compact | compact (-VOT) |
| | 16 msecs | [d-g] | compact | compact (-VOT) |
| | 32 msecs | /g/ | compact | compact |
| /gI/ | 4 msecs | [d-g] | compact | compact (-VOT) |
| | 8 msecs | [d-g] | compact | compact (-VOT) |
| | 16 msecs | /g/ | compact | compact (-VOT) |
| /ge/ | 4 msecs | [d-g] | compact | compact (-VOT) |
| | 8 msecs | [d-g] | compact | compact (-VOT) |
| | 16 msecs | /g/ | compact | compact |
| /gɛ/ | 4 msecs | [d-g] | compact | compact (-VOT) |
| | 8 msecs | /g/ | compact | compact |

(-VOT) no voicing onset; third or fourth frame used for 'spectrum at voicing onset'

* voicing onset at consonant release

shifts. The templates were fit both to the onset spectral shape, as well as to the spectral shape at onset of voicing. In cases where there was no onset of voicing, the spectrum at the fourth frame [10] was used as the spectral shape at 'voicing onset'. For some of the stimuli, particularly /b/'s, voicing onset was at consonant release, so the same spectrum was used for both 'onset' spectrum and 'spectrum at voicing onset'.

It is interesting to note that in most of the cases for Speakers PFA and MLD (8 out of 8 for Speaker PFA and 7 out of 9 for Speaker MLD), the templates appropriately fit the spectral shapes *at voicing onset* for the unambiguous, longer stimuli. However, the templates often did not fit the appropriate *stimulus onset* spectral shapes, as outlined in the previous section.

For Speaker PFA, five out of eight ambiguous tokens fit the two 'appropriate' templates at the third or fourth frame when there was *no voicing onset* for the very short stimuli; as more of the stimulus was appended, the spectral shape *at voicing onset* was appropriate to the changed dominant response. Thus, for this speaker only, response shifts could perhaps be interpreted as a response to the change in spectral shape *before voicing onset, though not at stimulus onset* vs. spectral shape *at voicing onset*. This would be a very interesting aspect to test in further experiments.

---

[10] When the amplitude of the spectrum was clearly too low at the fourth frame (i.e., peak values did not reach 30 dB), the spectrum at the third frame was used as the spectral shape at 'voicing onset'.

Tables 5.14-5.17 show the Lahiri and Blumstein ratios for each of the stimuli which had reponse shifts. These ratios require 'spectrum at voicing onset'; however, for some of these brief stimuli, particularly /d/'s and /g/'s, voicing was not present. For these cases, the spectrum at the third or fourth frame was used. Results indicate that the Lahiri and Blumstein ratios do not generally account for the category response shifts; for example, of a total of 13 cases where there was a responses shift from ambiguous [b-d] to /b/, only two cases show the appropriate change in the ratio values for the shorter stimuli as compared to the longer stimuli.

Tables 5.18-5.21 show the results of applying Kewley-Port's features to these stimuli. Results indicate that these features also do not generally account for the category response shifts. Of a total of 24 cases where a response shift was obtained as more of the signal was appended, only two cases have different Kewley-Port features for the longer stimulus compared to the shorter stimulus, and in neither case is this difference *appropriate* to the response shift.

Tables 5.22-5.25 show a summary and comparison of the Stevens and Blumstein templates, the Lahiri and Blumstein ratios and Kewley-Port's features. Results show that of the three spectral shape feature sets, the

### TABLE 5.14

### Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject PFA

| Syllable | Stimulus Length | Dominant Response | Ratio Value |
|---|---|---|---|
| /bI/ | 4 msecs | [b-d] | -.91 (-VOT) |
| | 8 msecs | /b/ | -2.1 |
| /be/ | 4 msecs | [b-d] | 2.5 (-VOT) |
| | 8 msecs | [b-d] | .5 |
| | 16 msecs | /b/ | -0.1 |
| /bɛ/ | 4 msecs | [b-d] | 2.5 (-VOT) |
| | 8 msecs | [b-d] | 6.0 |
| | 16 msecs | [b-d] | .42 |
| | 32 msecs | [b-d] | .54 |
| /bo/ | 4 msecs | [b-d] | 1.0 (-VOT) |
| | 8 msecs | /b/ | -4.0 |
| /bU/ | 4 msecs | [b-d] | .84 (-VOT) |
| | 8 msecs | /b/ | 2.4 |
| /dU/ | 4 msecs | [d-g] | .12 (-VOT) |
| | 8 msecs | [d-g] | .47 (-VPT) |
| | 16 msecs | /d/ | .08 |
| /du/ | 4 msecs | [d-g] | 1.0 (-VOT) |
| | 8 msecs | /d/ | -.33 |
| /dɚ/ | 4 msecs | [d-g] | .52 (-VOT) |
| | 8 msecs | [d-g] | 0.0 (-VOT) |
| | 16 msecs | /d/ | -.17 |

(-VOT) indicates no voicing onset; the fourth frame was used for 'spectrum at voicing onset'

## TABLE 5.15

## Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject RAH

| Syllable | Stimulus Length | Dominant Response | Ratio Value |
|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | -3.75 * |
| | 8 msecs | /b/ | 2.5 * |
| /bI/ | 4 msecs | [b-d-g] | 2.3 |
| | 8 msecs | /b/ | .78 |
| /be/ | 4 msecs | [b-d] | .6 * |
| | 8 msecs | [b-d] | .34 * |
| | 16 msecs | [b-d] | .5 * |
| | 32 msecs | [b-d] | .30 * |
| /bɛ/ | 4 msecs | [b-d-g] | 1.33 * |
| | 8 msecs | [b-d] | 2.33 * |
| | 16 msecs | [b-d] | .78 * |
| | 32 msecs | [b-d] | .91 * |
| /bæ/ | 4 msecs | [b-d] | .33 * |
| | 8 msecs | [b-d] | .64 * |
| | 16 msecs | [b-d] | .55 * |
| | 32 msecs | [b-d] | .76 * |
| /gɚ/ | 4 msecs | [d-g] | .71 (-VOT) |
| | 8 msecs | /g/ | 1.67 (-VOT) |

(-VOT) indicates no voicing onset; the last frame was used for 'spectrum at voicing onset'.

* The fourth frame was used for 'spectrum at voicing onset' since voicing occurred at consonant release.

## TABLE 5.16

### Natural Data With Response Shifts; Lahiri and Blumstein Features;

### *Subject* TMD

| Syllable | Stimulus Length | Dominant Response | Ratio Value |
|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | 1.33 |
| | 8 msecs | [b-d-g] | 4.0 |
| | 16 msecs | /b/ | 1.33 |
| /bI/ | 4 msecs | [b-d] | 1.2 |
| | 8 msecs | /b/ | 14.0 |
| /be/ | 4 msecs | [b-d] | .5 |
| | 8 msecs | /b/ | 2.5 |
| /bɛ/ | 4 msecs | [b-d] | .71 |
| | 8 msecs | /b/ | -8.0 |
| /dI/ | 4 msecs | [d-g] | -1.0 (-VOT) |
| | 8 msecs | [d-g] | -.22 (-VOT) |
| | 16 msecs | /d/ | 1.0 |
| /dæ/ | 4 msecs | [d-g] | 4.0 (-VOT) |
| | 8 msecs | [d-g] | 4.0 (-VOT) |
| | 16 msecs | /d/ | .64 |

(-VOT): no voicing onset; the fourth spectrum was used for 'spectrum at voicing onset'

## TABLE 5.17

## Natural Data With Response Shifts; Lahiri and Blumstein Features; Subject MLD

| Syllable | Stimulus Length | Dominant Response | Ratio Value |
|---|---|---|---|
| /be/ | 4 msecs | [b-d] | -.5 * |
| | 8 msecs | /b/ | 3.0 * |
| /bɛ/ | 4 msecs | [b-d] | .80 |
| | 8 msecs | [b-d] | .44 |
| | 16 msecs | /b/ | .71 |
| /bʌ/ | 4 msecs | [b-d] | 1.25 ** |
| | 8 msecs | [b-d] | 1.29 * |
| | 16 msecs | /b/ | .08 * |
| /bu/ | 4 msecs | [b-d-g] | .82 (-VOT) |
| | 8 msecs | [b-d-g] | .45 (-VOT) |
| | 16 msecs | [b-d] | .38 |
| | 32 msecs | [b-d-g] | .75 |
| /dæ/ | 4 msecs | [d-g] | .25 (-VOT) |
| | 8 msecs | [d-g] | .66 |
| | 16 msecs | /d/ | .33 |
| /gi/ | 4 msecs | [d-g] | .6 (-VOT) |
| | 8 msecs | [d-g] | 1.67 (-VOT) |
| | 16 msecs | [d-g] | .88 (-VOT) |
| | 32 msecs | /g/ | 1.08 |
| /gI/ | 4 msecs | [d-g] | .5 (-VOT) |
| | 8 msecs | [d-g] | 3.0 (-VOT) |
| | 16 msecs | /g/ | -1.0 (-VOT) |
| /ge/ | 4 msecs | [d-g] | -1.33 (-VOT) |
| | 8 msecs | [d-g] | -1.33 (-VOT) |
| | 16 msecs | /g/ | -.6 |
| /gɛ/ | 4 msecs | [d-g] | .63 (-VOT) |
| | 8 msecs | /g/ | .44 |

## TABLE 5.18

### Natural Data with Response Shifts; Kewley-Port Features; Subject PFA

| Syllable | Stimulus Length | Dominant Response | Features |
|---|---|---|---|
| /bI/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /be/ | 4 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | /b/ | rising spectrum |
| /bƐ/ | 4 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| | 8 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| | 16 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| | 32 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| /bo/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /bu/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /dU/ | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | [d-g] | rising spectrum |
| | 16 msecs | /d/ | rising spectrum |
| /du/ | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | /d/ | rising spectrum |
| /dɚ/ | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | [d-g] | rising spectrum |
| | 16 msecs | /d/ | rising spectrum |

## TABLE 5.19

## Natural Data with Response Shifts; Kewley-Port Features; Subject RAH

| Syllable | Stimulus Length | Dominant Response | Features |
|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /bI/ | 4 msecs | [b-d-g] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /be/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | [b-d] | rising spectrum |
| | 32 msecs | [b-d] | rising spectrum |
| /bɛ/ | 4 msecs | [b-d-g] | rising spectrum |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | [b-d] | rising spectrum |
| | 32 msecs | [b-d] | rising spectrum |
| /bæ/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | [b-d] | rising spectrum |
| | 32 msecs | [b-d] | rising spectrum |
| /gɚ/ | 4 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 8 msecs | /g/ | rising spectrum, mid-frequency peaks |

## TABLE 5.20

## Natural Data with Response Shifts; Kewley-Port Features; Subject TMD

| Syllable | Stimulus Length | Dominant Response | Features |
|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | falling spectrum |
| | 8 msecs | [b-d-g] | falling spectrum |
| | 16 msecs | /b/ | falling spectrum |
| /bI/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /be/ | 4 msecs | [b-d] | falling spectrum |
| | 8 msecs | /b/ | falling spectrum |
| /bɛ/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | /b/ | rising spectrum |
| /dI/ | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | [d-g] | rising spectrum |
| | 16 msecs | /d/ | rising spectrum |
| /dæ/ | 4 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 8 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 16 msecs | /d/ | rising spectrum, mid-frequency peaks |

## TABLE 5.21

### Natural Data with Response Shifts; Kewley-Port Features; Subject MLD

| Syllable | Stimulus Length | Dominant Response | Features |
|---|---|---|---|
| /be/ | 4 msecs | [b-d] | rising spectrum, mid-frequency peaks |
| | 8 msecs | /b/ | rising spectrum |
| /bɛ/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | /b/ | rising spectrum |
| /bʌ/ | 4 msecs | [b-d] | rising spectrum |
| | 8 msecs | [b-d] | rising spectrum |
| | 16 msecs | /b/ | rising spectrum |
| /bu/ | 4 msecs | [b-d-g] | rising spectrum |
| | 8 msecs | [b-d-g] | rising spectrum |
| | 16 msecs | [b-d] | rising spectrum |
| | 32 msecs | [b-d-g] | rising spectrum |
| /d / | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | [d-g] | rising spectrum |
| | 16 msecs | /d/ | rising spectrum |
| /gi/ | 4 msecs | [d-g] | rising spectrum |
| | 8 msecs | [d-g] | rising spectrum |
| | 16 msecs | [d-g] | rising spectrum |
| | 32 msecs | /g/ | rising spectrum |
| /gɪ/ | 4 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 8 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 16 msecs | /g/ | rising spectrum, mid-frequency peaks |
| /ge/ | 4 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 8 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 16 msecs | /g/ | rising spectrum, mid-frequency peaks |
| /gɛ/ | 4 msecs | [d-g] | rising spectrum, mid-frequency peaks |
| | 8 msecs | /g/ | rising spectrum, mid-frequency peaks |

## TABLE 5.22

## Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject PFA

| Syllable | Stimulus Length | Obtained Dominant Response | Predicted Response | | | |
|---|---|---|---|---|---|---|
| | | | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) | Lahiri & Blumstein | Kewley-Port |
| /bI/ | 4 msecs | [b-d] | /d/ | /d/ | /d/ | /d/ |
| | 8 msecs | /b/ | /d/ | /b/ | /d/ | /d/ |
| /be/ | 4 msecs | [b-d] | /d/ | [b-d] | /b/ | [d-g] |
| | 8 msecs | [b-d] | /d/ | [b-d] | [b-d] | /d/ |
| | 16 msecs | /b/ | /d/ | /b/ | /d/ | /d/ |
| /bɛ/ | 4 msecs | [b-d] | /d/ | [b-d] | /b/ | [d-g] |
| | 8 msecs | [b-d] | /d/ | [b-d] | /b/ | [d-g] |
| | 16 msecs | [b-d] | /d/ | [b-d] | /d/ | [d-g] |
| | 32 msecs | [b-d] | /d/ | [b-d] | [b-d] | [d-g] |
| /bo/ | 4 msecs | [b-d] | /d/ | /d/ | /b/ | /d/ |
| | 8 msecs | /b/ | /d/ | /b/ | /d/ | /d/ |
| /bU/ | 4 msecs | [b-d] | [b-d] | [b-d] | /b/ | /d/ |
| | 8 msecs | /b/ | [b-d] | /b/ | /b/ | /d/ |
| /dU/ | 4 msecs | [d-g] | /g/ | [d-g] | /d/ | /d/ |
| | 8 msecs | [d-g] | /g/ | [d-g] | [b-d] | /d/ |
| | 16 msecs | /d/ | /g/ | /d/ | /d/ | /d/ |
| /du/ | 4 msecs | [d-g] | /g/ | [d-g] | /b/ | /d/ |
| | 8 msecs | /d/ | /g/ | /d/ | /d/ | /d/ |
| /dɚ/ | 4 msecs | [d-g] | /g/ | [d-g] | [b-d] | /d/ |
| | 8 msecs | [d-g] | /g/ | [d-g] | /d/ | /d/ |
| | 16 msecs | /d/ | /g/ | /d/ | /d/ | /d/ |

## TABLE 5.23

## Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject RAH

| Syllable | Stimulus Length | Obtained Dominant Response | Predicted Response | | | |
|---|---|---|---|---|---|---|
| | | | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) | Lahiri & Blumstein | Kewley-Port |
| /bi/ | 4 msecs | [b-d-g] | /d/ | /d/ | /d/ | /d/ |
| | 8 msecs | /b/ | /d/ | /d/ | /b/ | /d/ |
| /bI/ | 4 msecs | [b-d-g] | /d/ | /d/ | /b/ | /d/ |
| | 8 msecs | /b/ | /d/ | /d/ | /b/ | /d/ |
| /be/ | 4 msecs | [b-d] | /d/ | /d/ | /b/ | /d/ |
| | 8 msecs | [b-d] | /d/ | /d/ | /d/ | /d/ |
| | 16 msecs | [b-d] | /d/ | /d/ | [b-d] | /d/ |
| | 32 msecs | [b-d] | /d/ | /d/ | /d/ | /d/ |
| /bɛ/ | 4 msecs | [b-d-g] | [b-d] | [b-d] | /b/ | /d/ |
| | 8 msecs | [b-d] | /d/ | [b-d] | /b/ | /d/ |
| | 16 msecs | [b-d] | /d/ | [b-d] | /b/ | /d/ |
| | 32 msecs | [b-d] | /d/ | [b-d] | /b/ | /d/ |
| /bæ/ | 4 msecs | [b-d] | [b-g] | [b-g] | /d/ | /d/ |
| | 8 msecs | [b-d] | /b/ | [b-d] | /b/ | /d/ |
| | 16 msecs | [b-d] | /b/ | [b-d] | [b-d] | /d/ |
| | 32 msecs | [b-d] | /b/ | [b-d] | /b/ | /d/ |
| /gɚ/ | 4 msecs | [d-g] | [b-g] | [b-g] | /b/ | [d-g] |
| | 8 msecs | /g/ | [b-g] | [b-g] | /b/ | [d-g/ |

## TABLE 5.24

## Natural Data With Response Shifts;

## Summary and Comparison of Spectral Paramater Sets;

## Subject TMD

| Syllable | Stimulus Length | Obtained Dominant Response | Predicted Response | | | |
|---|---|---|---|---|---|---|
| | | | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) | Lahiri & Blumstein | Kewley-Port |
| /bi/ | 4 msecs | [b-d-g] | [b-g] | [b-g] | /b/ | /b/ |
| | 8 msecs | [b-d-g] | [b-g] | [b-g] | /b/ | /b/ |
| | 16 msecs | /b/ | [b-g] | [b-g] | /b/ | /b/ |
| /bI/ | 4 msecs | [b-d] | [b-d] | [b-d] | /b/ | /d/ |
| | 8 msecs | /b/ | [b-d] | /b/ | /b/ | /d/ |
| /be/ | 4 msecs | [b-d] | [b-d-g] | [d-g] | [b-d] | /b/ |
| | 8 msecs | /b/ | [b-d-g] | /b/ | /b/ | /b/ |
| /bɛ/ | 4 msecs | [b-d] | /d/ | /d/ | /b/ | /d/ |
| | 8 msecs | /b/ | /d/ | /b/ | /d/ | /d/ |
| /dI/ | 4 msecs | [d-g] | [d-g] | [b-g] | /d/ | /d/ |
| | 8 msecs | [d-g] | [d-g] | [b-g] | /d/ | /d/ |
| | 16 msecs | /d/ | [d-g] | /d/ | /b/ | /d/ |
| /dæ/ | 4 msecs | [d-g] | /g/ | /g/ | /b/ | [d-g] |
| | 8 msecs | [d-g] | /g/ | /g/ | /b/ | [d-g] |
| | 16 msecs | /d/ | /g/ | /b/ | /b/ | [d-g] |

### TABLE 5.25

### Natural Data With Response Shifts; Summary and Comparison of Spectral Paramater Sets; Subject MLD

| Syllable | Stimulus Length | Obtained Dominant Response | Predicted Response | | | |
|---|---|---|---|---|---|---|
| | | | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) | Lahiri & Blumstein | Kewley-Port |
| /be/ | 4 msecs | [b-d] | [b-d-g] | [b-d-g] | /d/ | [d-g] |
| | 8 msecs | /b/ | [b-d-g] | [b-d-g] | /b/ | /d/ |
| /bɛ/ | 4 msecs | [b-d] | /d/ | /d/ | /b/ | /d/ |
| | 8 msecs | [b-d] | /d/ | /d/ | /d/ | /d/ |
| | 16 msecs | /b/ | /d/ | /b/ | /b/ | /d/ |
| /bʌ/ | 4 msecs | [b-d] | [b-d] | [b-d] | /b/ | /d/ |
| | 8 msecs | [b-d] | [b-d] | [b-d] | /b/ | /d/ |
| | 16 msecs | /b/ | [b-d] | /b/ | /d/ | /d/ |
| /bu/ | 4 msecs | [b-d-g] | /d/ | [d-g] | /b/ | /d/ |
| | 8 msecs | [b-d-g] | /d/ | [d-g] | [b-d] | /d/ |
| | 16 msecs | [b-d] | /d/ | [d-g] | /d/ | /d/ |
| | 32 msecs | [b-d-g] | /d/ | [d-g] | /b/ | /d/ |
| /dæ/ | 4 msecs | [d-g] | /d/ | /d/ | /d/ | /d/ |
| | 8 msecs | [d-g] | /d/ | /d/ | /b/ | /d/ |
| | 16 msecs | /d/ | /d/ | /d/ | /d/ | /d/ |
| /gi/ | 4 msecs | [d-g] | /g/ | /g/ | /b/ | /d/ |
| | 8 msecs | [d-g] | /g/ | /g/ | /b/ | /d/ |
| | 16 msecs | [d-g] | /g/ | /g/ | /b/ | /d/ |
| | 32 msecs | /g/ | /g/ | /g/ | /b/ | /d/ |
| /gI/ | 4 msecs | [d-g] | /g/ | /g/ | [b-d] | [d-g] |
| | 8 msecs | [d-g] | /g/ | /g/ | /b/ | [d-g] |
| | 16 msecs | /g/ | /g/ | /g/ | /d/ | [d-g] |
| /ge/ | 4 msecs | [d-g] | /g/ | /g/ | /d/ | [d-g] |
| | 8 msecs | [d-g] | /g/ | /g/ | /d/ | [d-g] |
| | 16 msecs | /g/ | /g/ | /g/ | /d/ | [d-g] |
| /gɛ/ | 4 msecs | [d-g] | /g/ | /g/ | /b/ | [d-g] |
| | 8 msecs | /g/ | /g/ | /g/ | [b-d] | [d-g] |

Stevens and Blumstein templates fit to the spectra at voicing onset account for more of the obtained dominant responses than the other feature sets. Of 81 cases, the Stevens and Blumstein templates fit 40 of the spectra at voicing onset (49%), but only 11 of the spectra at stimulus onset (14%); the Lahiri and Blumstein ratios are correct for only 16 of the 81 stimuli (20%), and the Kewley-Port features are correct for only 15 of them (19%). This is an indication that more complex templates (such as those given by Stevens and Blumstein) are required, which actually give some indication of the nature of the spectral shape, as opposed to simple rising and falling parameters.

#### 5.2.3.2 Analysis of Formant Information

Formant measurements for the stimuli which had response shifts are given in Tables 5.26-5.29. Formants were well tracked at these short stimulus durations for /b/'s and most /d/'s, but many /g/'s had fluctuating formant estimates, indicating the presence of longer bursts in /g/'s.

Tables 5.30-5.33 show the predicted values of F2i and F3i from the male and female regression lines for the two male and female speakers, respectively. Only those stimuli which showed a shift in responses category when 24 msec was presented (i.e., 16 msec full amplitude) were analyzed. Formant transitions were clearly evident for these stimuli. Predictions of F2i

## TABLE 5.26

## Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements;

### Subject PFA

| Syllable | Stimulus Length | Dominant Response | F2 | A2 | F3 | A3 |
|---|---|---|---|---|---|---|
| /bI/ | 4 msecs | [b-d] | 1830 | 31 | 2450 | 29 |
| | 8 msecs | /b/ | 1830 | 34 | 2340 | 32 |
| /be/ | 4 msecs | [b-d] | 1780 | 26 | 2400 | 26 |
| | 8 msecs | [b-d] | 1780 | 26 | 2460 | 27 |
| | 16 msecs | /b/ | 1780 | 32 | 2530 | 32 |
| /bɛ/ | 4 msecs | [b-d] | 1720 | 31 | 2520 | 30 |
| | 8 msecs | [b-d] | 1720 | 31 | 2580 | 31 |
| | 16 msecs | [b-d] | 1770 | 35 | 2580 | 32 |
| | 32 msecs | [b-d] | 1840 | 30 | 2640 | 30 |
| /bo/ | 4 msecs | [b-d] | 1100 | 33 | 2400 | 17 |
| | 8 msecs | /b/ | 1100 | 36 | 2460 | 21 |
| /bU/ | 4 msecs | [b-d] | 1030 | 28 | 2200 | 21 |
| | 8 msecs | /b/ | 1100 | 32 | 2330 | 21 |
| /dU/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | 1720 | 18 | 2460 | 19 |
| | 16 msecs | /d/ | 1650 | 30 | 2340 | 31 |
| /du/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | /d/ | 1590 | 12 | 2340 | 21 |
| /dɚ/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | 1590 | 21 | 2400 | 22 |
| | 16 msecs | /d/ | 1530 | 34 | 2330 | 31 |

- fluctuating formant track (burst position)

## TABLE 5.27

### Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements;

### Subject RAH

| Syllable | Stimulus Length | Dominant Response | F2 | A2 | F3 | A3 |
|---|---|---|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | 2080 | 27 | 2640 | 33 |
| | 8 msecs | /b/ | 2150 | 27 | 2700 | 33 |
| /bI/ | 4 msecs | [b-d-g] | 2030 | 29 | 2700 | 30 |
| | 8 msecs | /b/ | 2030 | 29 | 2710 | 31 |
| /be/ | 4 msecs | [b-d] | 1900 | 29 | 2650 | 33 |
| | 8 msecs | [b-d] | 1960 | 31 | 2700 | 34 |
| | 16 msecs | [b-d] | 1960 | 32 | 2700 | 36 |
| | 32 msecs | [b-d] | 1970 | 30 | 2760 | 32 |
| /bɛ/ | 4 msecs | [b-d-g] | 1900 | 29 | 2640 | 33 |
| | 8 msecs | [b-d] | 1960 | 32 | 2710 | 32 |
| | 16 msecs | [b-d] | 1960 | 30 | 2710 | 32 |
| | 32 msecs | [b-d] | 1960 | 30 | 2710 | 30 |
| /bæ/ | 4 msecs | [b-d] | 1720 | 31 | 2570 | 31 |
| | 8 msecs | [b-d] | 1720 | 32 | 2580 | 34 |
| | 16 msecs | [b-d] | 1770 | 34 | 2580 | 33 |
| | 32 msecs | [b-d] | 1830 | 32 | 2630 | 32 |
| /gɝ/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | /g/ | - | - | - | - |

- fluctuating formant track (burst position)

## TABLE 5.28

## Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements;

### Subject TMD

| Syllable | Stimulus Length | Dominant Response | F2 | A2 | F3 | A3 |
|---|---|---|---|---|---|---|
| /bi/ | 4 msecs | [b-d-g] | 2330 | 26 | 2830 | 29 |
| | 8 msecs | [b-d-g] | 2390 | 27 | 2880 | 30 |
| | 16 msecs | /b/ | 2460 | 24 | 2950 | 28 |
| /bI/ | 4 msecs | [b-d] | 2080 | 28 | 2650 | 25 |
| | 8 msecs | /b/ | 2140 | 29 | 2700 | 25 |
| /be/ | 4 msecs | [b-d] | 2020 | 30 | 2580 | 25 |
| | 8 msecs | /b/ | 2030 | 32 | 2580 | 25 |
| /bɛ/ | 4 msecs | [b-d] | 1970 | 32 | 2640 | 26 |
| | 8 msecs | /b/ | 1960 | 30 | 2640 | 26 |
| /dI/ | 4 msecs | [d-g] | 2220 | 19 | 2770 | 23 |
| | 8 msecs | [d-g] | 2230 | 27 | 2710 | 26 |
| | 16 msecs | /d/ | 2330 | 27 | 2780 | 30 |
| /dæ/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | 2080 | 32 | 2760 | 29 |
| | 16 msecs | /d/ | 2150 | 31 | 2770 | 30 |

- fluctuation formant track (burst position)

## TABLE 5.29

### Natural Data With Response Shifts; Formant (F) and Formant Amplitude (A) Measurements;

### Subject MLD

| Syllable | Stimulus Length | Dominant Response | F2 | A2 | F3 | A3 |
|---|---|---|---|---|---|---|
| /be/ | 4 msecs | [b-d] | - | - | - | - |
| | 8 msecs | /b/ | 2210 | 28 | 2830 | 21 |
| /bɛ/ | 4 msecs | [b-d] | 1900 | 19 | 2890 | 20 |
| | 8 msecs | [b-d] | 2080 | 20 | 2960 | 16 |
| | 16 msecs | /b/ | 2210 | 31 | 3010 | 28 |
| /bʌ/ | 4 msecs | [b-d] | - | - | - | - |
| | 8 msecs | [b-d] | 1400 | 35 | 2770 | 25 |
| | 16 msecs | /b/ | 1530 | 38 | 2940 | 30 |
| /bu/ | 4 msecs | [b-d-g] | - | - | - | - |
| | 8 msecs | [b-d-g] | - | - | - | - |
| | 16 msecs | [b-d] | 1540 | 36 | 2710 | 23 |
| | 32 msecs | [b-d-g] | 1580 | 35 | 2770 | 25 |
| /dæ/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | 2090 | 33 | 3080 | 29 |
| | 16 msecs | /d/ | 2030 | 35 | 3020 | 30 |
| /gi/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | - | - | - | - |
| | 16 msecs | [d-g] | - | - | - | - |
| | 32 msecs | /g/ | 2530 | 24 | 3070 | 26 |
| /gI/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | - | - | - | - |
| | 16 msecs | /g/ | - | - | - | - |
| /ge/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | [d-g] | - | - | - | - |
| | 16 msecs | /g/ | 2650 | 20 | 3140 | 25 |
| /gɛ/ | 4 msecs | [d-g] | - | - | - | - |
| | 8 msecs | /g/ | - | - | - | - |

- fluctuating formant track (burst position)

## TABLE 5.30

## Predicted Values of F2i and F3i From Male Formant Measurements; Natural Data With Response Shifts; Speaker PFA

| Syllable | Obtained Formant Values | | | | | Predicted Formant Values | |
|---|---|---|---|---|---|---|---|
| | F2i | F2f | F3i | F3f | | F2i | F3i |
| /be/ | 1780 | 1780 | 2400 | 2530 | b | 1682 | 2439 |
| | | | | | d | 1865 | 2627 |
| | | | | | g | 2007 | 2665 |
| /bɛ/ | 1720 | 1840 | 2520 | 2640 | b | 1733 | 2496 |
| | | | | | d | 1894 | 2689 |
| | | | | | g | 2070 | 2771 |
| /dU/ | 1720 | 1650 | 2460 | 2340 | b | 1572 | 2341 |
| | | | | | d | 1804 | 2519 |
| | | | | | g | 1870 | 2482 |
| /dɝ/ | 1590 | 1530 | 2400 | 2330 | b | 1470 | 2336 |
| | | | | | d | 1747 | 2514 |
| | | | | | g | 1745 | 2472 |

## TABLE 5.31

## Predicted Values of F2i and F3i From Male Formant Measurements; Natural Data With Response Shifts; Speaker RAH

| Syllable | Obtained Formant Values | | | | | Predicted Formant Values | |
|---|---|---|---|---|---|---|---|
| | F2i | F2f | F3i | F3f | | F2i | F3i |
| /be/ | 1900 | 1970 | 2650 | 2760 | b | 1844 | 2558 |
| | | | | | d | 1956 | 2757 |
| | | | | | g | 2207 | 2886 |
| /bɛ/ | 1900 | 1960 | 2640 | 2710 | b | 1835 | 2532 |
| | | | | | d | 1951 | 2729 |
| | | | | | g | 2196 | 2838 |
| /bæ/ | 1720 | 1830 | 2570 | 2630 | b | 1725 | 2491 |
| | | | | | d | 1889 | 2684 |
| | | | | | g | 2060 | 2761 |

## TABLE 5.32

## Predicted Values of F2i and F3i From Female Formant Measurements: Natural Data With Response Shifts; Speaker TMD

| Syllable | Obtained Formant Values | | | | | Predicted Formant Values | |
|---|---|---|---|---|---|---|---|
| | F2i | F2f | F3i | F3f | | F2i | F3i |
| /bi/ | 2330 | 2460 | 2830 | 2950 | b | 2262 | 2733 |
| | | | | | d | 2259 | 2915 |
| | | | | | g | 2699 | 3024 |
| /dİ/ | 2220 | 2330 | 2770 | 2780 | b | 2150 | 2642 |
| | | | | | d | 2208 | 2840 |
| | | | | | g | 2517 | 2835 |
| /dæ/ | 2080 | 2150 | 2760 | 2770 | b | 1996 | 2637 |
| | | | | | d | 2136 | 2836 |
| | | | | | g | 2330 | 2824 |

## TABLE 5.33

## Predicted Values of F2i and F3i From Female Formant Measurements: Natural Data With Response Shifts; Speaker MLD

| Syllable | Obtained Formant Values | | | | | Predicted Formant Values | |
|---|---|---|---|---|---|---|---|
| | F2i | F2f | F3i | F3f | | F2i | F3i |
| /bɛ/ | 1900 | 2210 | 2890 | 3010 | b | 2047 | 2765 |
| | | | | | d | 2160 | 2941 |
| | | | | | g | 2392 | 3090 |
| /bʌ/ | 1400 | 1530 | 2770 | 2940 | b | 1463 | 2728 |
| | | | | | d | 1891 | 2911 |
| | | | | | g | 1685 | 3012 |
| /bu/ | 1540 | 1580 | 2710 | 2770 | b | 1506 | 2637 |
| | | | | | d | 1911 | 2836 |
| | | | | | g | 1737 | 2824 |
| /dæ/ | 2090 | 2030 | 3080 | 3020 | b | 1893 | 2771 |
| | | | | | d | 2089 | 2946 |
| | | | | | g | 2205 | 3101 |

and F3i were based on the formant values at 16 msec; these values, denoted as F2f and F3f, were used as the F2v and F3v values.

Tables 5.34-5.37 show the deviations of the predicted F2i and F3i values from the actual formant measurements provided in Tables 5.30-5.33. The predicted categories are based on minimum deviation from predicted values and are shown in Tables 5.38 (male speakers) and 5.39 (female speakers). Tables 5.38 and 5.39 show predicted categories from the F2 regression lines, the F3 regression lines, and a combination of F2 and F3 information. This combined F2 and F3 category prediction is based on the absolute values of the sum of F2 and F3 deviations from predicted values. In addition, the predicted categories from Stevens and Blumstein templates fit at stimulus onset and voicing onset are provided for comparison.

Formant information seems to aid in disambiguating the signal; as more of the signal is appended, more transition information is provided. In 8 out of 14 cases (for both male and female speakers) where there was a response category shift after 24 msec of the signal was presented, this addition of formant information strengthened the correct category response as measured by minimum distance from the regression lines of Figure 2.8. However, the transition *slopes* (i.e., rising, falling or straight transtions) were appropriate in *all*

## TABLE 5.34

## Deviations of Predicted Values From Male Formant Measurements; Natural Data With Response Shifts; Speaker PFA

| Syllable | | Deviations | | | | |
|---|---|---|---|---|---|---|
| | | F2i | | F3i | | F2i and F3i |
| | | Hz | % | Hz | % | Hz |
| /be/ | b | 98 | 5.8 | -39 | 1.6 | 137 |
| | d | -85 | 4.5 | -227 | 8.6 | 312 |
| | g | -227 | 11.3 | -265 | 9.9 | 492 |
| /bɛ/ | b | -13 | 0.8 | 24 | 1.0 | 37 |
| | d | -174 | 9.2 | -169 | 6.3 | 343 |
| | g | -350 | 16.9 | -251 | 9.1 | 601 |
| /dU/ | b | 148 | 9.4 | 119 | 5.1 | 267 |
| | d | -84 | 4.7 | -59 | 2.3 | 143 |
| | g | -150 | 8.0 | -22 | 0.9 | 172 |
| /dɝ/ | b | 120 | 8.2 | 64 | 2.7 | 184 |
| | d | -157 | 9.0 | -114 | 4.5 | 271 |
| | g | -155 | 8.9 | -72 | 2.9 | 227 |

## TABLE 5.35

## Deviations of Predicted Values From Male Formant Measurements; Natural Data With Response Shifts; Speaker RAH

| Syllable | | Deviations | | | | |
|---|---|---|---|---|---|---|
| | | F2i | | F3i | | F2i and F3i |
| | | Hz | % | Hz | % | Hz |
| /be/ | b | 56 | 3.0 | 92 | 3.6 | 148 |
| | d | 56 | 2.9 | -107 | 3.9 | 163 |
| | g | -307 | 13.9 | -236 | 8.1 | 543 |
| /bɛ/ | b | 65 | 3.5 | 108 | 4.3 | 173 |
| | d | -51 | 2.6 | -89 | 3.3 | 140 |
| | g | -296 | 13.5 | -198 | 7.0 | 494 |
| /bæ/ | b | -5 | 0.3 | 79 | 3.2 | 84 |
| | d | -169 | 8.9 | -114 | 4.2 | -283 |
| | g | -340 | 16.5 | -191 | 6.9 | 531 |

## TABLE 5.36

## Deviations of Predicted Values From Female Formant Measurements; Natural Data With Response Shifts; Speaker TMD

| Syllable | | Deviations | | | | |
|---|---|---|---|---|---|---|
| | | F2i | | F3i | | F2i and F3i |
| | | Hz | % | Hz | % | Hz |
| /bi/ | b | 68 | 3.0 | 97 | 3.5 | 165 |
| | d | 71 | 3.1 | -85 | 2.9 | 156 |
| | g | -369 | 13.7 | -194 | 6.4 | 563 |
| /dI/ | b | 70 | 3.3 | 128 | 4.8 | 198 |
| | d | 12 | 0.5 | -70 | 2.5 | 82 |
| | g | -297 | 11.8 | -65 | 2.3 | 362 |
| /dæ/ | b | 84 | 4.2 | 123 | 4.7 | 207 |
| | d | -56 | 2.6 | -76 | 2.7 | 132 |
| | g | -250 | 10.7 | -64 | 2.3 | 314 |

## TABLE 5.37

## Deviations of Predicted Values From Female Formant Measurements; Natural Data With Response Shifts; Speaker MLD

| Syllable | | Deviations | | | | |
|---|---|---|---|---|---|---|
| | | F2i | | F3i | | F2i and F3i |
| | | Hz | % | Hz | % | Hz |
| /bɛ/ | b | -147 | 7.2 | 125 | 4.5 | 272 |
| | d | -260 | 12.0 | -51 | 1.7 | 311 |
| | g | -492 | 20.6 | -200 | 6.5 | 692 |
| /bʌ/ | b | -63 | 4.3 | 42 | 1.5 | 105 |
| | d | -491 | 26.0 | -141 | 4.8 | 632 |
| | g | -285 | 16.9 | -242 | 8.0 | 527 |
| /bu/ | b | 34 | 2.3 | 73 | 2.8 | 107 |
| | d | -371 | 19.4 | -126 | 4.4 | 497 |
| | g | -197 | 11.3 | -114 | 4.0 | 311 |
| /dæ/ | b | 197 | 10.4 | 309 | 11.1 | 506 |
| | d | 1 | 0.1 | 134 | 4.5 | 135 |
| | g | -115 | 5.2 | -21 | 0.7 | 136 |

## TABLE 5.38

## Summary and Comparison of Obtained and Predicted Final Response Categories; Natural Data With Response Shifts; Male Speakers

Speaker PFA

| Syllable | Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|---|
| | | F2i vs. F2v | F3i vs. F3v | F2 & F3 | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) |
| /be/ | /b/ | /d/ | /b/ | /b/ | /d/ | /b/ |
| /bɛ/ | [b-d] | /b/ | /b/ | /b/ | /d/ | [b-d] |
| /dU/ | /d/ | /d/ | /d/ | /d/ | /g/ | /d/ |
| /dɝ/ | /d/ | /b/ | /b/ | /b/ | /g/ | /d/ |

Speaker RAH

| Syllable | Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|---|
| | | F2i vs. F2v | F3i vs. F3v | F2 & F3 | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) |
| /be/ | [b-d] | [b-d] | /b/ | /b/ | /d/ | /d/ |
| /bɛ/ | [b-d] | /d/ | /d/ | /d/ | /d/ | [b-d] |
| /bæ/ | [b-d] | /b/ | /b/ | /b/ | /b/ | [b-d] |

## TABLE 5.39

## Summary and Comparison of Obtained and Predicted Final Response Categories; Natural Data With Response Shifts; Female Speakers

Speaker TMD

| Syllable | Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|---|
| | | F2i vs. F2v | F3i vs. F3v | F2 & F3 | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) |
| /bi/ | /b/ | /b/ | /d/ | /d/ | [b-g] | [b-g] |
| /dI/ | /d/ | /d/ | /g/ | /d/ | [d-g] | /d/ |
| /dæ/ | /d/ | /d/ | /g/ | /d/ | /g/ | /b/ |

Speaker MLD

| Syllable | Obtained Category | Predicted Categories | | | | |
|---|---|---|---|---|---|---|
| | | F2i vs. F2v | F3i vs. F3v | F2 & F3 | Stevens & Blumstein (onset) | Stevens & Blumstein (voicing) |
| /bɛ/ | /b/ | /b/ | /d/ | /b/ | /d/ | /b/ |
| /bʌ/ | /b/ | /b/ | /b/ | /b/ | [b-d] | /b/ |
| /bu/ | [b-d-g] | /b/ | /b/ | /b/ | /d/ | [d-g] |
| /dæ/ | /d/ | /d/ | /g/ | /d/ | /d/ | /d/ |

cases. This was particularly evident for /b/'s before front vowels, which normally have long transitions (cf. Dorman, et al., 1977). Appending more of the /b/ signal naturally included addition of formant information; in all cases where the correct categorization of the signal was strengthened, formants were rising (as is appropriate for /b/'s). Tokens which obtained response shifts from ambiguous [d-g] to /d/ also showed that adding more of the formant transitions could be influencing subjects category responses; when more of the signal was appended, formants showed falling transitions for /d/'s before back vowels, and straight or even slightly rising transitions for /d/'s before front vowels. These are the same /d/ formant trends that were found in Cooper et al.'s (1952) formant perception study, in the measurement study in Chapter 2 (i.e, the regression lines of Fig. 2.8) and in the perception study of Chapter 3. This perhaps indicates that a formant *slope* measurment is required, or that the actual F2 steady-state, as opposed to the F2f that was used here, is needed. Thus, it is not unreasonable to suppose that the addition of formant information could be helpful in disambiguating these short stimuli.

However, the spectral shape *at voicing onset* is also appropriate for many of the final majority responses (10 out of the 14 cases). The role of spectral shape at voicing onset merits further investigation.

However, it should also be noted that this factor seems to show speaker differences (refer to discussion above regarding Speaker PFA). It would also not explain the results of the synthetic experiment of Chapter 3, in which voicing excitation was sythesized at stimulus onset.

The results of this study indicate that formant information plays a role in natural speech as well as in the perception studies which have used synthetic speech. Results also coincide with Dorman et al.'s (1977) suggestion of 'cue weighting': if longer transitions are apparent (e.g., as in the case of /b/ transitions) then transitions are a 'stronger' cue than the burst. In addition, work by Pols and Schouten (1985) indicate that transitions, but not bursts, are more salient in sentence-context than in isolation; this would indicate that for speech perception in general (in addition to speech processing in highly constrained experiments such as these), transitions play a very important role.

### 5.2.4 Summary

Results of these studies indicate that both spectral and formant information seem to be important in the perception of natural speech stimuli, as well as in perception of synthetic stimuli (which was tested in Chapter 3). In particular, onset spectral shape seems to account more for /g/ responses, while formant information seems to account more for /b/ and /d/ responses. In addition, when a /d/ 'locus' is present, this also appears to cue a /d/.

These studies are a presentation of preliminary analysis of these stimuli. Further investigation is required as to the exact nature of the role of formant and spectral shape interaction. Of particular interest is the notion of analyzing the spectral shape characterisitics before voicing onset (but not at stimulus onset) and at voicing onset.

In addition, several proposals for refining the spectral template shapes were given, and merit further investigation. However, since formant informantion seems to be an important cue, subsequent template revisions should perhaps incorporate this information for those consonant categories which were most affected by formant transitions (i.e., /b/ and /d/).

Another very important refinement should be the automatization of template fitting. Automatic template fitting and categorization would provide a good test of the effectiveness of the templates.

# 6. SUMMARY AND CONCLUSIONS

The present study has investigated the role of formant onsets in various vowel contexts for the perception of stop consonants. Spectral shape characteristics were also analyzed. Perception studies using both natural and synthetic speech showed that both formant and spectral shape information were important factors.

## 6.1 The Role of Formant Onsets

The measurement study of Chapter 2 showed consistent linear patterns when formant onsets were plotted against vowel steady-states for both male and female speech. Measurements of F2 were in accord with early perception studies (Cooper et al., 1952).

Subsequent perception studies were done on the vowels ranging from /o/-/U/-/ɛ/ using cascade-synthesized speech. Results were similar to results obtained in early perception studies which had used a *different* synthesis technique and *different* vowel contexts (Cooper et al., 1952; Hoffman, 1958; Harris et al., 1958). This indicates that the results of these studies using synthetic speech are not merely reflecting ad hoc perceptual strategies for specific types of synthetic stimuli.

In general, subjects response categorizations showed that, for vowels with low F2 steady-states, low and mid F2i's were heard as /g/ and higher F2i's as /d/; however, for vowels with higher steady-states, low F2i's were heard as /b/, mid F2i's as /d/ and high F2i's as /g/. The '/d/-/g/ crossover' in the perception study closely matched the '/d/-/g/ crossover' in the measurement of natural data.

Linear trends were also found in natural data measurement for F3i vs. F3v; categorization data from the perception study can also be related to these F3 measurements. These results were in accord with the early perception studies of Hoffman (1958) and Harris et al. (1958).

Category predictions were calculated based on minimum distance from the regression lines of Figure 2.8. In general, F2i vs. F2v made better predictions than F3i vs. F3v. Predictions based on F2i and F3i combined made the same predictions as F2i vs. F2v; this indicates the relative strength of the predicitons made by F2i vs. F2v and suggests that the F2 regression lines provide a 'stronger' cue (see Dorman et al., 1977).

In general, F2i predictions for all consonant categories were correct for higher F2v's. At lower F2v's, /d/ predictions were correct but stimuli which were predicted to be /b/ were categorized by listeners as /g/.

Results of the perception study using natural speech indicated that response shifts could also be related to predictions based on the regression lines of Figure 2.8.

## 6.2 The Role of Spectral Shape

Spectral shapes of both synthetic and natural stimuli were analyzed using

1. Stevens and Blumstein templates,
2. Lahiri and Blumstein ratios, and
3. Kewley-Port features.

In general, the Stevens and Blumstein templates made better predictions than the Lahiri and Blumstein ratios and Kewley-Port features. However, automatic template and feature extraction techniques are desirable.

Results showed that the shape of the spectrum at stimulus onset (for both natural and synthetic stimuli) could not account for subjects' responses. Stimuli categorized as /g/ *did* tend to have 'compact' spectral shapes but the compact templates also falsely accepted many stimuli that were categorized as /b/'s or /d/'s by listeners. However, an analysis of the spectral shape at *voicing onset*, for natural data only, tended to have the appropriate compact shape for /g/, the diffuse rising shape for /d/ and the diffuse falling shape for /b/. Further analysis of this factor would be very interesting.

Specific proposals were given for the redefinition of spectral shape features. These included frequency peak locations, slope of the rise and fall from peak location, and some new interpretations for the feature 'compact' (see Chapter 3). In addition, a suggestion of hierarchical template-fitting was implemented and shown to decrease the number of false alarms.

## 6.3 The Role of the Vowel Label

The results of the perception study presented in Chapter 4 show that the role of labelling the vowel context is minimal and inconsistent. When the choice of the vowel label is stringently controlled, subject differences emerged regarding the effect of the vowel label on consonant categorization. Response patterns by the *majority* of subjects were not significantly altered when the vowel context was labelled as different vowels (even though the same acoustic information had been presented). Only one subject showed an effect of the vowel label on her consonant categorizations, but her responses did not seem to be interpretable on the basis of formant based 'syllable-coding' (see Lieberman, 1984).

## 6.4 The Effect of Vowel Context

Results show that vowel context is an important factor in both the production and perception of stop consonants. Vowel context affected stop consonant categorizations for both synthetic and natural stimuli. These results are contrary to claims of acoustic 'invariance' over vowel categories (e.g., Cole and Scott, 1974; Rao, 1974).

Vowel context was found to be important in formant measurements (see Figure 2.8), as well as in the spectral shape analysis that was done for natural speech tokens. For example, natural tokens of /d/ before front vowels often had a peak value near the 'locus' (Delattre et. al., 1955), which was not the case with /d/'s in a back or central vowel context . In addition, natural tokens of /g/ before back or central vowels were more likely to have two prominent peaks, while /g/'s in a front vowel context showed only one prominent peak.

Vowel context affected stop consonant categorization using both synthetic and natural speech stimuli. The response patterns for a synthetic continuum of formant onsets in a low F2v context were different than when they were synthesized in a high F2v context (see Figure 3.3). Identifications of short natural stimuli showed that /b/'s before front vowels were not as well identified as /b/'s before back vowels.

The identification of /b/'s in general, for both synthetic and natural speech stimuli, seems to be a rather complex matter. Both formant and spectral parameters predicted majority /b/ responses for low F2i's in a low F2v context; however, subjects consistently categorized these stimuli as /g/'s (see Chapter 3 for possible explanations). The identification of brief tokens of natural /b/'s showed that /b/'s in a front vowel context, which has a *higher* F2v, were more poorly identified than /b/'s in a back or central context, which have *lower* F2v's. It is precisely the /b/'s in front vowel contexts which 'require' more tranisition information (i.e., obtained majority /b/ responses only after a longer portion of the signal was presented). Thus, perhaps formant transitions are more important for /b/'s in front vowel contexts than in back or central vowel contexts. Further study of this question is desirable.

In general, both spectral and formant information made similar predictions of consonant categorizations for consonants followed by vowels with high F2v's. For low F2v contexts, however, stimuli having a compact spectral shape were often categorized as /g/'s, while /d/ categorizations seemed to be based on formant criteria. Thus, vowel context affects formant and spectral shape predictions.

### 6.5 Suggestions for Future Reseach

There are a number of related topics based on the results of this study which merit further investigation. They include cross-lnguistic studies, manipulation of other vowel properties, experiments using filtered speech, and automatic consonant categorization.

A cross-linguistic study of the roles of formant and spectral information would be very interesting. Do other languages show similar formant and spectral properties of stop consonants (Fant, 1973; Fischer-Jorgenson, 1954; Blumstein and Lahiri, 1982)? Would listeners of other language groups show the same general response patterns as English listeners? Of particular interest is the cross-linguistic validation of the 'd-g crossovers' that were found in both perception and natural data measurment in

English. Finally, what is the role of vowel context labelling in other languages, particularly for languages with a smaller number of vowels?

Studies could also be done in which other acoustic properties could be manipulated in various vowel contexts. For example, the F3 steady-state could be altered (Shammass and Nearey, in preparation), or dynamic properties such as diphthongization or rates of transition could be investigated. Other acoustic properties of interest include burst amplitude (see Repp, 1978) and F1.

A series of experiments using filtered speech is recommended, as well. High pass filtering would tend to raise the spectral tilt, while low pass filtering would lower the spectral tilt. Specific questions regarding spectral shape could then be addressed, such as the proposed feature of spectral shape at voicing onset.

Finally, automatic implementation of template fitting, as well as automatic estimation of formant deviations from the regression lines, should be encouraged. Automatic classification of stop consonants could provide a very strong test of correct feature selection. Furthermore, more complex models of the relationship between spectral shape and formant frequency transitions are necessary.

The problem of invariance in stop consonants is wide-ranging and highly topical. This research project attempted to shed some light on the role of vowel context, but there are many more interesting questions which could be investigated. The above suggestions merely give a few possible avenues for future research.

# REFERENCES

Assmann, P.F. *1979*. The role of context in vowel perception. Unpublished M.Sc. thesis, Department of Linguistics, University of Alberta.

Bell, C.G., Fujisaki H., Heinz, J.M., Stevens K.N., and House, A.S. *1961*. Reduction of speech by analysis-by-synthesis techniques. J. Acoust. Soc. Am. 33: 1725-1736.

Blumstein, S.E., Stevens, K.N., and Nigro G.N. *1977*. Property detectors for bursts and transitions in speech perception. J. Acoust. Soc. Am. 61: 1301-1313.

Blumstein, S.E., and Stevens, K.N. *1979*. Acoustic invariance in speech production: evidence from meassurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Am. 66: 1001-1017.

Blumstein, S.E. and Stevens, K.N. *1980*. Perceptual invariance in onset spectra for stop consonants in different vowel environments. J. Acoust. Soc. Am. 67: 648-662.

Blumstein, S.E., Issacs. E., and Mertus, J. *1982*. The role of the gross spectral shape as a perceptual cue to place of articulation in in initial stop consonants. J. Acout. Soc. Am. 72: 43-50.

Bock, R.D. *1975*. Multivariate statistical methods in behavioral research. McGraw Hill: New York.

Bridle J.S. and Sedgwick, N.C. *1977*. A method for segmenting acoustic patterns with applications to automatic speech recognition. IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25: 656-659.

Chistovitch, L.A., Kozhevnikov,V.A., Lesogor, L.W., Shuplijakov, V.S., Taljasin, P.A. and Tjulkov, W.A. *1974*. A functional model of signal processing in the peripheral auditory system. Acustica 31: 349-353.

Cole, R.A. and Scott, B. *1974*. The phantom in the phoneme: invariant cues for stop consonants. Perc. and Psychophys. 15: 101-107.

Cooper, F.S., Delattre P.C., Liberman, A.M., BorstJ.M., and Gerstman, L.J., *1952*. Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Am. 24: 597-606.

Cooper, W.E. *1974*. Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Am. 56: 617-627.

Christensen, R., Strong, W. and Palmer, E. 1976. A comparison of three methods of extracting resonance information from predictor-coefficient coded speech. IEEE Trans. Acoust. Speech and Signal Processing Vol. ASSP-24: 8-14.

Demichelis, P., DeMori, R., Laface, P., and O'Kane, M. 1979. Computer recognition of stop consonants. IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27: 85-88.

Delattre, P.C., Liberman, A.M., Cooper, F.S.1955. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27: 769-773.

Denes, P.B. 1967. On the motor theory of speech perception. In Models for the Perception of Speech and Visual Form. Proceedings of a Symposium, 1964. W. Walter-Dunn, ed. MIT Press: Cambridge, MA.

Dolmazon, J.M. Bastet, L., and Shupljakov,V.S. 1977. A functional model of the peripheral auditory system in speech processing. IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25: 261-264.

Dorman, M.F. and Studdert-Kennedy, M. and Raphael, L.J. 1977. Stop consonant recognition: release bursts and formant transition as functionally equivalent context dependent cues. Percept. Psych. 22: 109-122.

Edwards, T.J. 1981. Multiple feature analysis of intervocalic English plosives. J. Acoust. Soc. Am. 69: 535-547.

Eimas, P.D., Siqueland, E.R., Jusczyk, P. and Vigorito, J. 1971. Speech perception in infants. Science 171: 303-306.

Fant, G. 1960. Acoustic theory of speech production. Mouton:The Hague.

Fant, G. 1973. Stops in CV syllables. In Speech Sounds and Features G. Fant, ed. MIT Press: Cambridge, MA. 110-139.

Fienberg, S.E. 1981. The analysis of cross-classified categorical data. MIT Press: Cambridge, MA.

Fischer-Jorgensen, E. 1954. Acoustic analysis of stop consonants. Miscellanea Phonetica, 2: 42-59.

Fry, D.B., Abramson, A.S, Eimas, P.D. and Liberman, A.M. 1962. The identification and discrimination of synthetic vowels. *Lang. and Speech 5:* 171.

Fujimura, O. *1974*. The syllable as a unit of speech recognition. *IEEE Symp. Speech Recognition* April, 1974.

Fujisaki, H. and Tominaga, M. *1982*. Automatic recognition of voiced stop consonants in CV and VCV utterances. *IEEE Acoust., Speech, Signal Processing, Vol. ASSP-30:* 1996-1999.

Graham, C.H. *1966*. Visual perception. In *Handbook of Experimental Psychology* S.S. Stevens, ed. John Wiley and Sons Inc.: New York. 868-920.

Haberman, J.H. *1979*. *Analysis of qualitative data: Vol. I and II.* Academic Press: New York.

Halle, M., Hughes, G. and Radley, J. *1957*. Acoustic properties of stop consonants. *J. Acoust. Soc. Am. 29:* 107-116.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C. and Cooper, F.S. *1958*. Effect of third formant transitions on the perception of the voiced stop consonants. *J. Acoust. Soc. Am. 30:* 122-126.

Hochberg, J.E. *1964*. *Perception*. Prentice-Hall Inc.: New Jersey.

Hoffman, H.S. *1958*. Study of some cues in the perception of the voiced stop consonants. *J. Acoust. Soc. Am. 30:* 1035-1041.

Hubel, D. and Wiesel, T. *1962*. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol. 160:* 106-154.

Jacobsen, R., Fant, C.G.M. and Halle, M. *1952*. *Preliminaries to Speech Analysis*. MIT Press: Cambridge, MA.

Just, M.A., Suslick R.L., Michaels, S., and Schockey, S. *1978*. Acoustic cues and phychological processes in the perception of natural stop consonants. *Perc. and Psychophys. 24:* 327-336.

Kewley-Port, D. *1980*. Respresentations of spectral change as cues to place of articulation in stop consonants. *Res. Speech Percept: Tech. Rep. No. 3* Department of Psychology, Indiana Universtity.

Klatt, D.H. *1980*. Speech perception: a model of acoustic-phonetic analysis and lexical access. In *Perception and Production of Fluent Speech* R.A. Cole, ed. Lawrence Erlbaum Associates: New Jersey.

Klatt, D.H. *1980*. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am. 67:* 971-995.

Kobatake, H. and Noso, K. *1980*. Feature extraction and recognition of Japanese voiceless stop consonants by principal component analysis. *J. Acoust. Soc. Jpn (E)1:* 215-227.

Koenig, W., Dunn, H.K., and Lacy, L.Y. *1946*. The sound spectrograph. *J. Acoust. Soc. Am. 17:* 19-49.

Kuhn, G.M. *1979*. Stop consonant place perception with single-formant stimuli: evidence for the role of the front cavity resonance. *J. Acoust. Soc. Am. 65:* 774-788.

Lahiri, A. and Blumstein, S.E. S.E. *1981*. A reconsideration of acoustic invariance for place of articulation in stop consonants: evidence from cross-language studies. Paper presented at the 102 meeting of Acoust. Soc. of Am., Miami Beach, Florida, Nov. 30-Dec. 4, 1981.

Lennig, M. *1978*. Acoustic measurement of linguistic change: the modern Paris vowel system. Ph.D. Dissertation, Pennsylvania Dissertation Series: No. 1, Philadelphia, PA.

Lieberman, P. *1979*. Phonetics and physiology: some current issues.In *Perspectives in Experimental Linguistics* .G.D. Prideaux, ed., John Benjamins: Amsterdam.

Mariani, J.J. and Lienard, J.S. *1977*. Acoustic-phonemic recognition of connected speech using transient information. *IEEE Trans Acoust., Speech, Signal Processing, Vol. ASSP-25:* 667-670.

Markel, J.D. and Grey, A.H. *1976*. *Linear prediction of speech*. Springer-Verlag: Berlin.

Massaro, D.W. and Oden, G.C. *1980*. Evaluation and integration of acoustic features in speech perception. *J. Acoust. Soc. Am. 67:* 996-1013.

Massaro, D.W. and Cohen, M.M. *1983*. Phonological context in speech perception. *Percept. and Psychophys. 34:* 338-348.

Menon, K.M., Rao, P.V., and Thorsar, R.B. *1974*. Formant transitions and stop consonant perception in syllables. *Lang. and Speech 19:* 27-46.

Mermelstein, P. *1978*. On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perc. and Psychophys. 23:* 331-336.

Molho, L. *1976*. Automatic recognition of fricatives and plosives in continuous speech.". IEEE Acoust., Speech, Signal Processing, Vol. ASSP-24: 182-185.

Nearey, T.M. *1977*. Phonetic feature systems for vowels. PhD. Dissertation, University of Conneticut. Reprinted by the Indiana University Linguistics Club, Indiana, 1978.

Nearey, T.M., Hogan, J.T. and Roszypal, A.J. *1979*. Speech signals, cues and features. In Perspectives in Experimental Linguistics. Current Issues in Linguistic Theory, Vol. 10. G.D. Prideaux, ed. John Benjamins B.V.: Amsterdam, 73-96.

Nearey, T.M. and Hogan, J.T. *in press*. Phonological contrast in experimental linguistics: relating distributions of measurements of natural data to categorization curves. In Experimental Phonology. J. Ohala, ed. Academic Press: New York.

Ohman, S. *1966*. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. 39: 151-168.

Paul, A.P., House, A.S., Stevens, K.N. *1964*. Automatic reduction of vowel spectra: an analysis by synthesis method and its evaluation. J. Acoust. Soc. Am. 36: 303-308.

Pols, L.C. and Schouten, M.E. *1985*. Plosive consonant identification in ambiguous sentences. J. Acoust. Soc. Am. 78: 33-39.

Rao, P.V. *1974*. On stop consonants. Speech Communication Seminar, Stockholm, Aug. 1974, 95-109.

Rabiner, L.R. and Schafer, R.W. *1978*. Digital processing of speech signals. Prentice-Hall: Englewood Cliffs.

Sawusch, S.R. *1977*. Processing of place information in stop consonants. Perc. and Psychophys. 22: 417-426.

Sawusch, J. and Pisoni, D.B. "On the Identification of *1974*. On the identification of place and voicing features in synthetic stop consonants. J. of Phonetics 2: 181-194.

Schatz, C.D. *1954*. The role of context in the perception of stops. Language 30: 47-56.

Schwartz, R., Klovstad, J., Makhoul, J., Sorenson, J. *1980*. A preliminary design of a phonetic vocoder based on a diphone model. IEEE Acoust., Speech, Signal Processing Vol. ASSP-28: 32-35.

Searle, C.L., Jacobson, J.Z. and Rayment, S.G. *1979*. Stop consonant discrimination based on human audition. J. Acoust. Soc. Am. 65: 799-809.

Shammass, S.E. and Nearey, T.M. (in preparation). Formant transitions as partial invariants in the identification of voiced stops.

Silverman, H.F. and Dixon, N.R. *1976*. The 1976 modular acoustic processor (MAP): diadic segment classification and final phonemic string estimation. IEEE Acoust., Speech, Signal Processing, Vol. ASSP-24: 15-20.

SPSS Inc. *1983*. SPSS-X users guide. McGraw-Hill: New York.

Stevens, K.N. and House, A.S. *1955*. Development of a quantitative description of vowel articulation. J. Acoust. Soc. Am. 27: 484-493.

Stevens, K.N. and Blumstein, S.E. *1978*. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Am. 64: 1358-1368.

Stevenson, D. and Stephens, R. *1979*. The Alligator Reference Manual. Unpublished manuscript.

Studdert-Kennedy, M. *1976*. Speech perception. In Contemporary Issues in Experimental Phonetics. N.J. Lass, ed. Academic Press: New York, 243-293.

Tekieli, M.E. and Cullinan, W.L. *1979*. The perception of temporally segmented vowels and consonant-vowel syllables. J. Speech Hear. Res. 22: 103-121.

Thompson, C. and Hollien, H. *1970*. Some contextual effects on the perception of synthetic vowels. Lang. and Speech 13: 1-13.

Tohkura, Y., Itakura, F. and Hashimoto, S. *1978*. Spectral smoothing technique in PARCOR speech analysis-synthesis. IEEE Trans. Acoust., Speech, Signal Processing Vol. ASSP-26: 587-596.

Velleman, P. and Hoaglin, D. *1981*. ABC's of EDA. Duxbury Press: New York.

Walley, A.C., and Carol, T.D. *1983*. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. J. Acoust. Soc. Am. 73: 1011-1022.

Wickelgren, W.A. *1969*. Context sensitive coding, associative memory and serial order in (speech) behavior". Psychol. Review: 76 1-15.

Willis, C. 1971. Synthetic vowel categorization and dielectology. Lang. and Speech 14: 213-228.

Winitz, H, Scheib, M.E. and Reeds, J.A. 1972. Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. J. Acoust. Soc. Am. 51: 1309-1317.

Wolf, C.G. 1978. Perceptual invariance for stop consonants in different positions. Perc. and Psychophys. 24: 315-326.

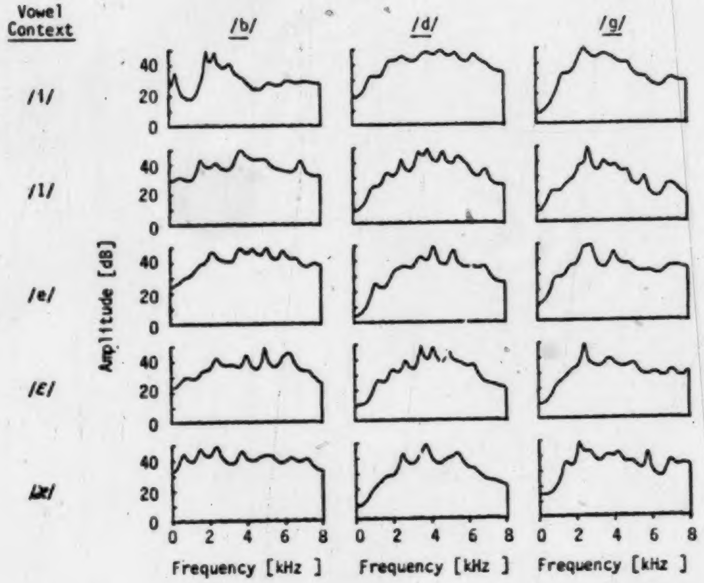
Weinstein, C.J., McCandless, S.S., Mondshein, L.F. and Zue, V.W. 1974. IEEE Trans. Acoust., Speech and Signal Processing Vol. ASSP-22: 54-67.

Zwicker, E., Terhardt, E. and Paulus, E. 1979. Automatic speech recognition using psychoaacoustic models. J. Acoust. Soc. Am. 65: 487-498.

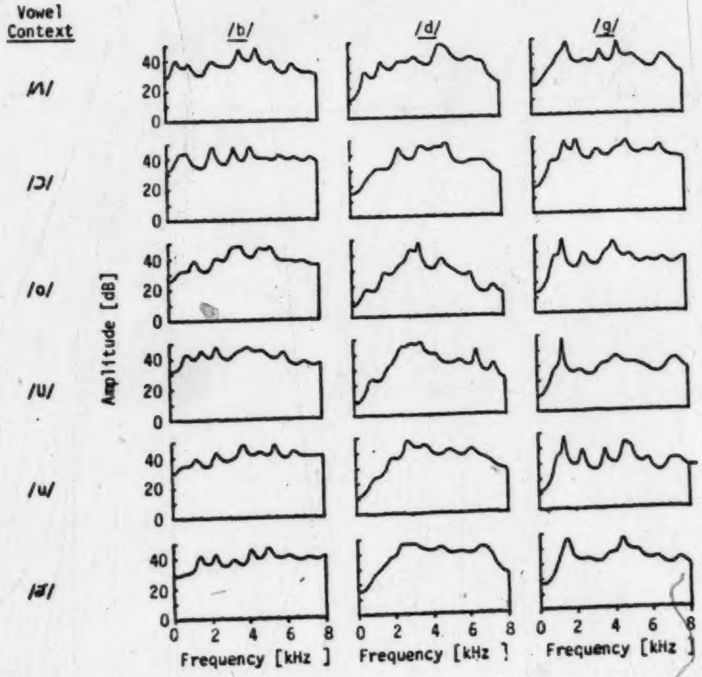
# APPENDIX I

## Spectra of Natural Data;

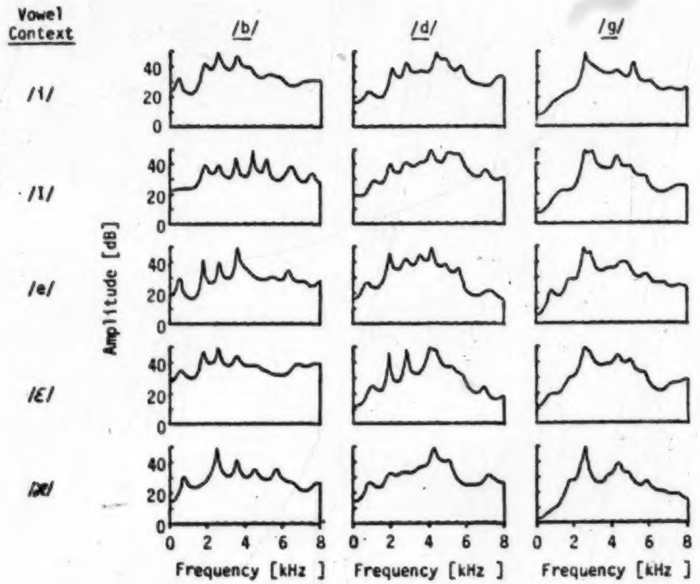
## 4 msec onset



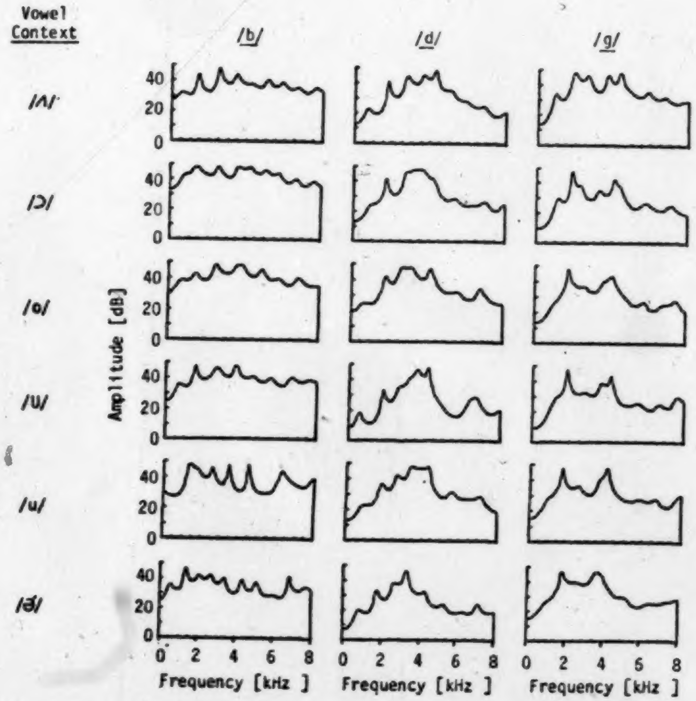
Front Vowel Context; Speaker PFA



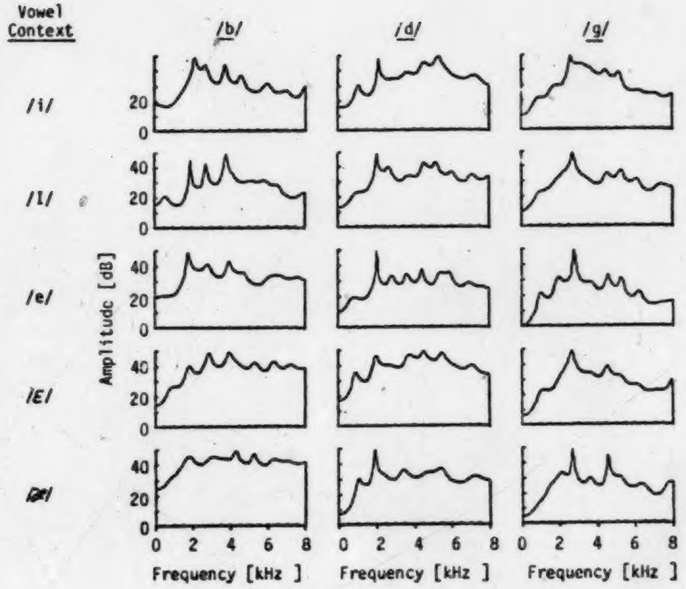
Central or Back Vowel Context; Speaker PFA



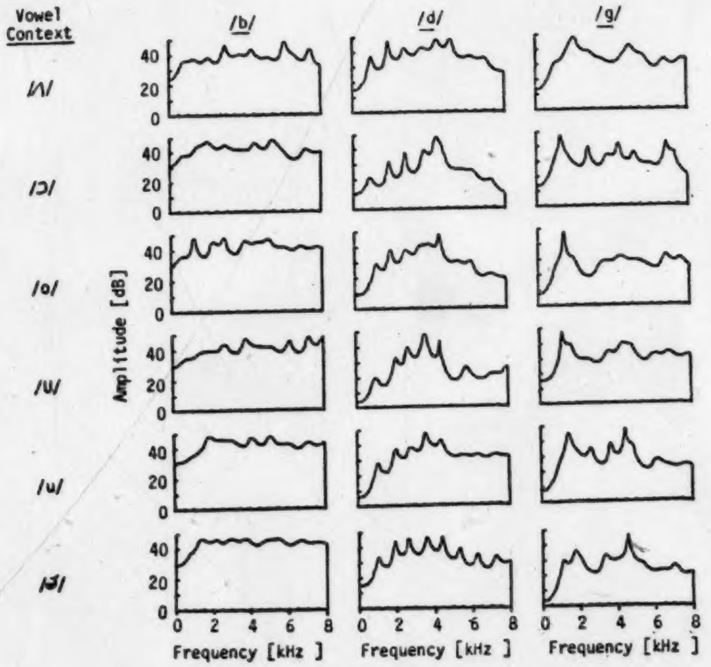
Front Vowel Context; Speaker RAH



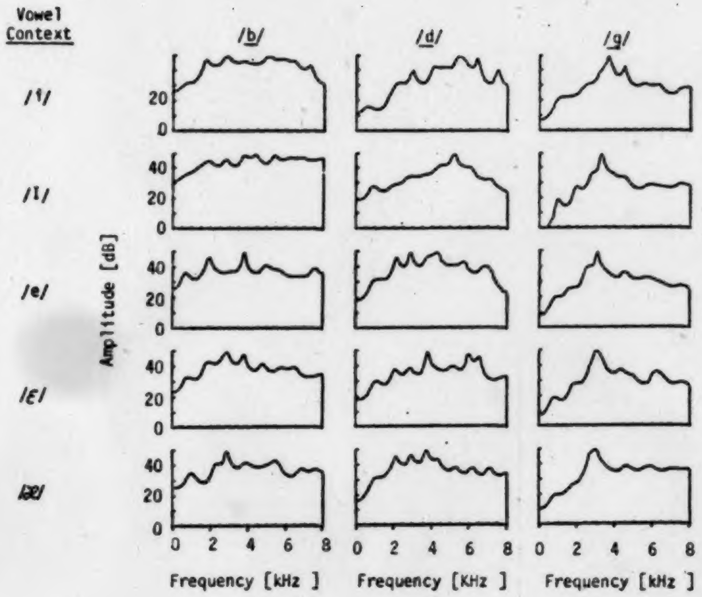
Central or Back Vowel Context; Speaker RAH



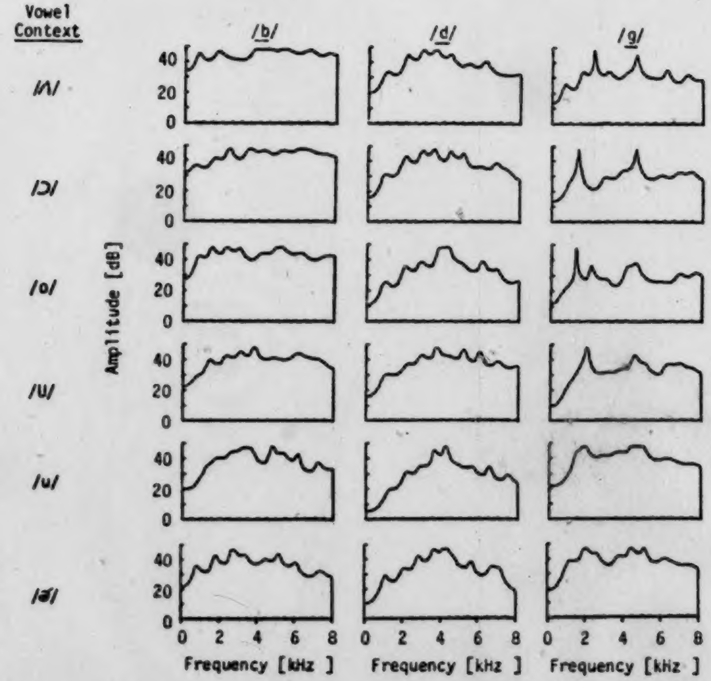
Front Vowel Context; Speaker TMD



Central or Back Vowel Context; Speaker TMD



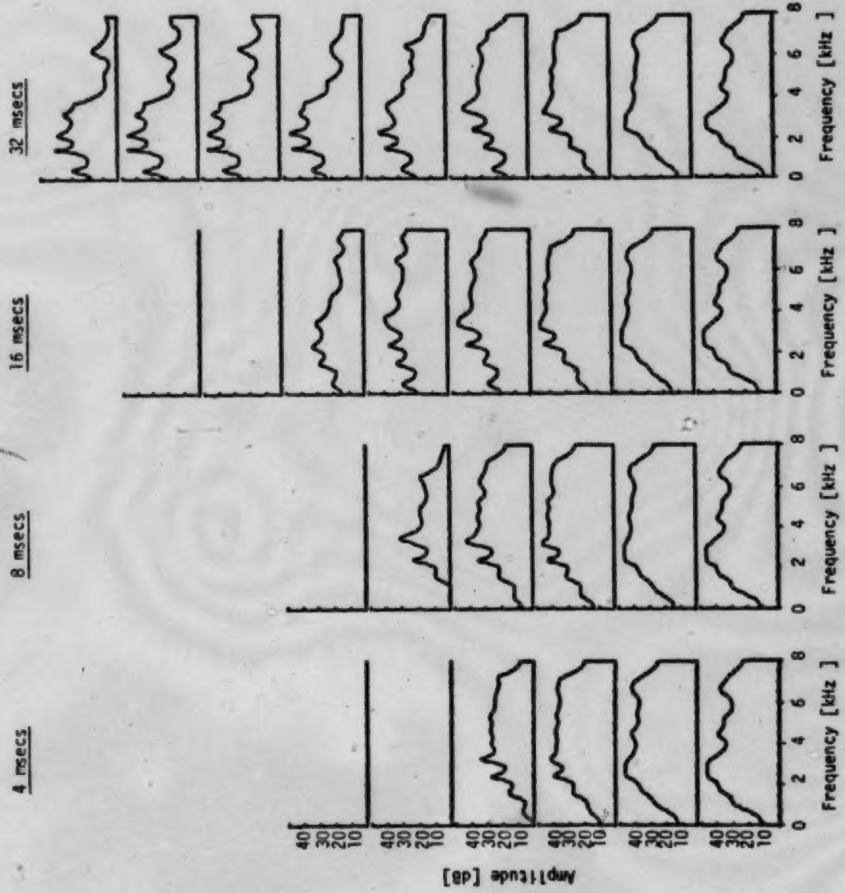
Front Vowel Context; Speaker MLD



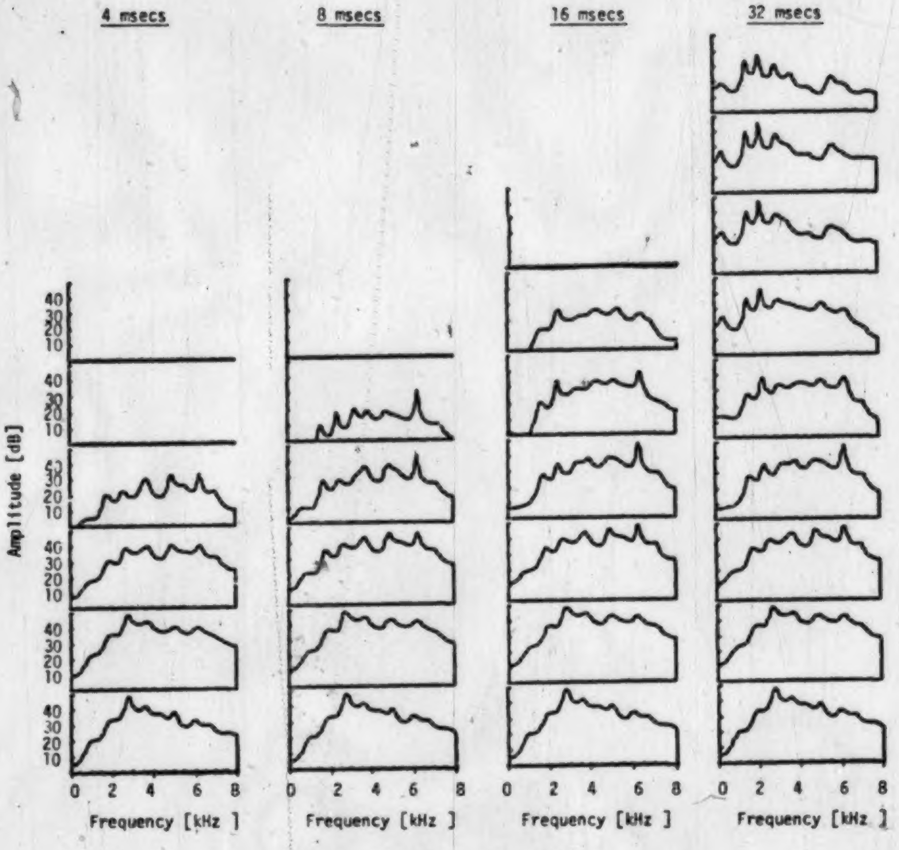
Central or Back Vowel Context; Speaker MLD

# APPENDIX II

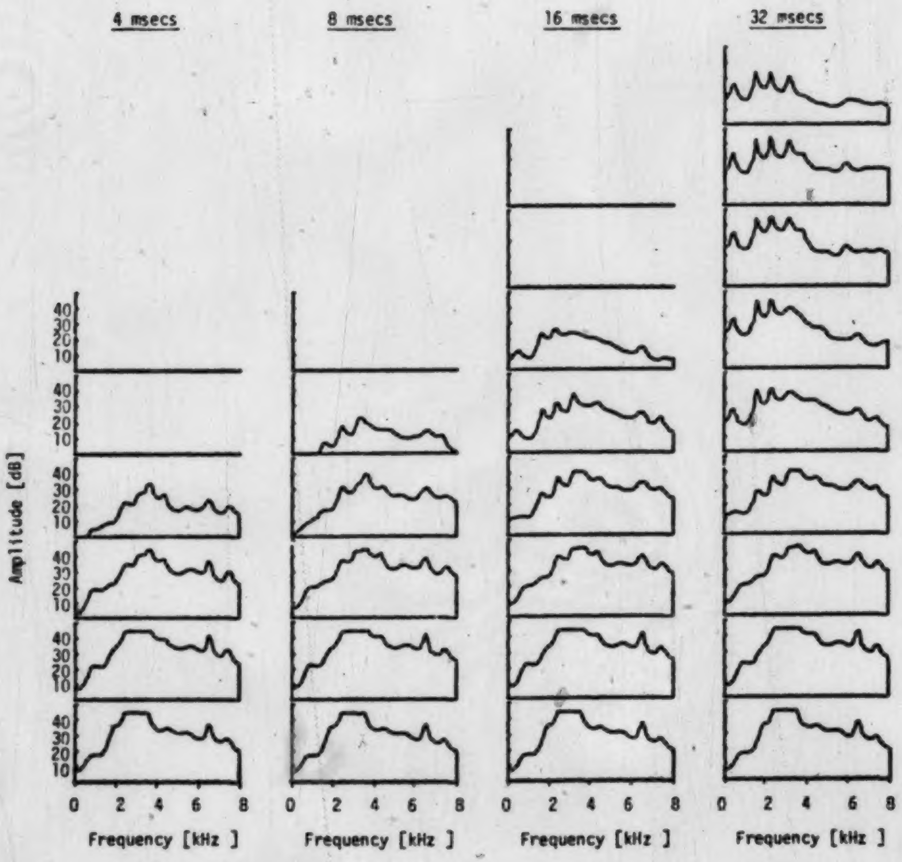
## Spectra of Natural Data With Response Shifts



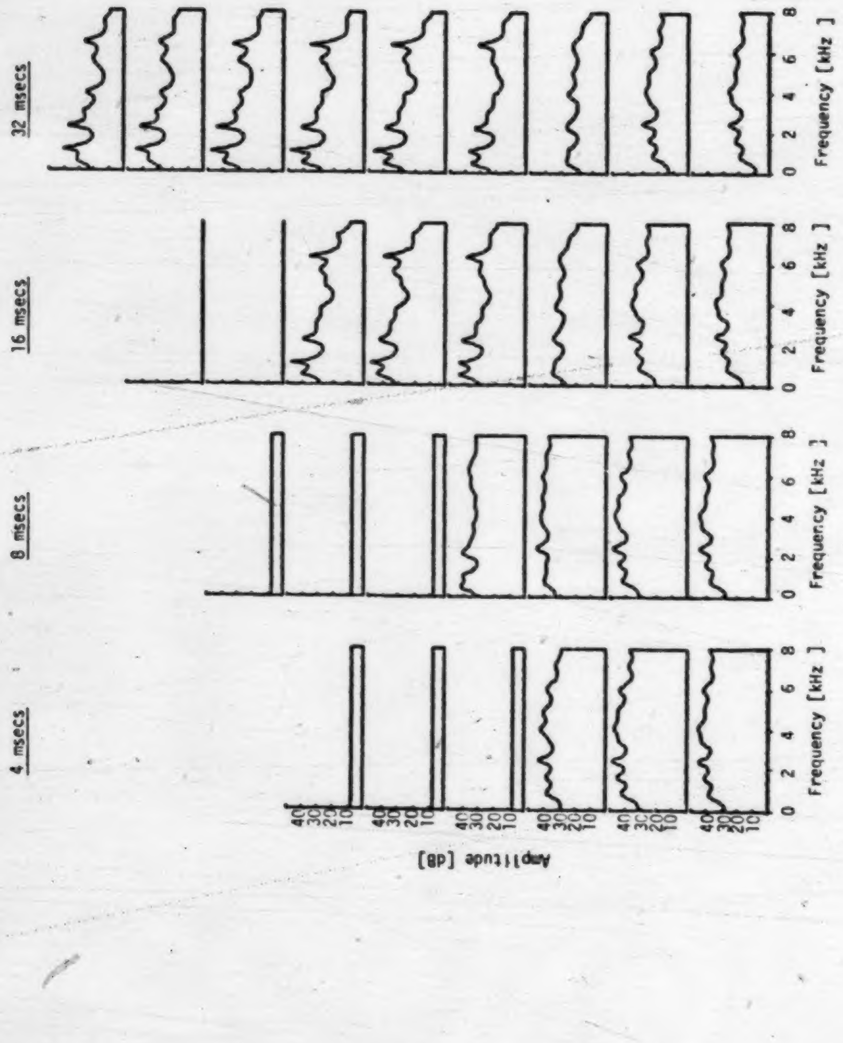
Syllable /dɚ/: Speaker PFA



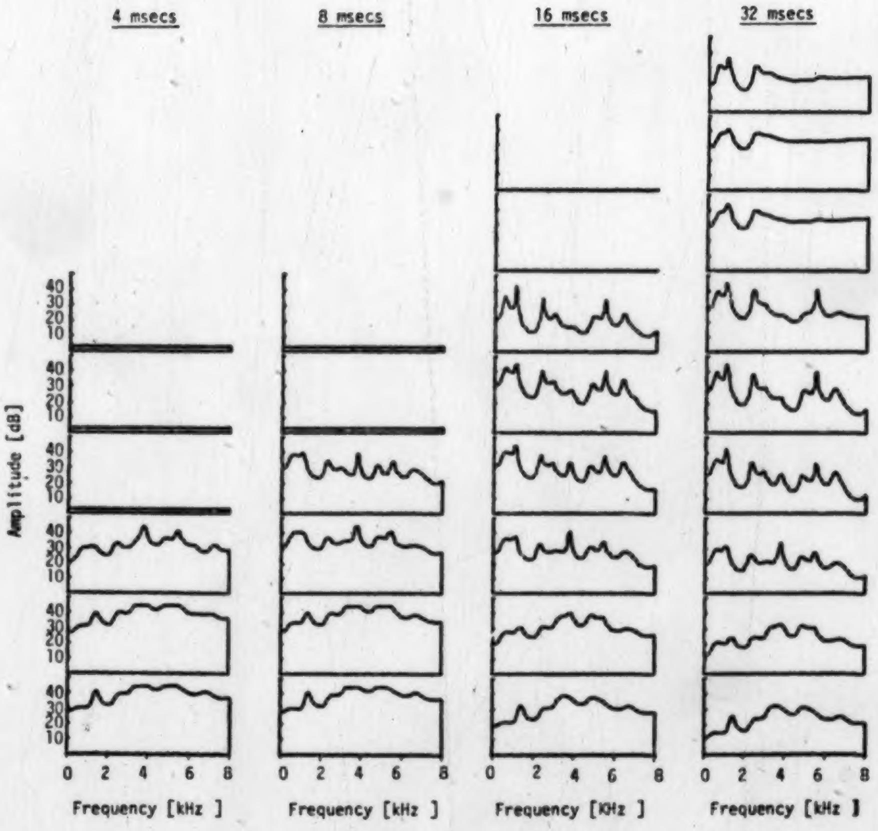
Syllable /du/; Speaker PFA



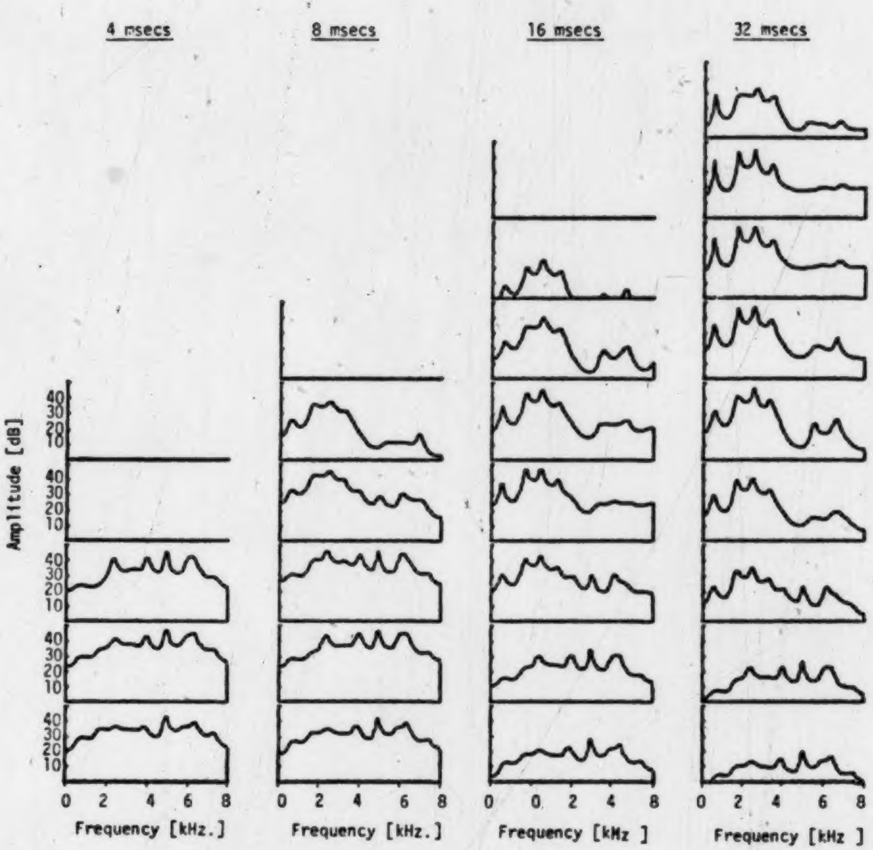
Syllable /dU/; Speaker PFA



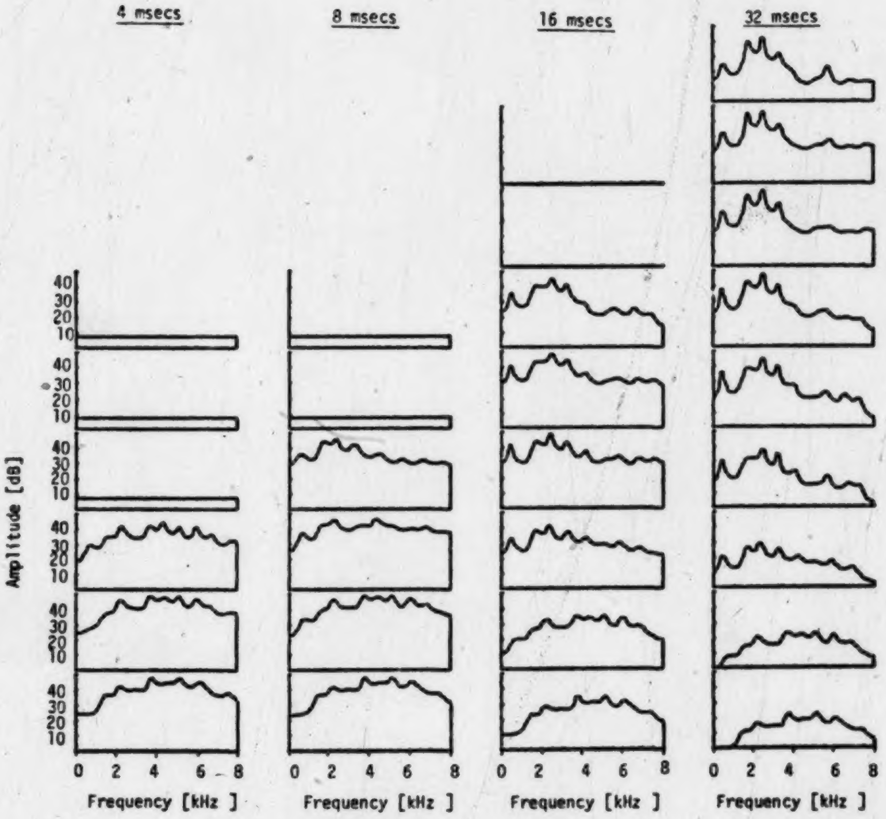
Syllable /bU/; Speaker PFA



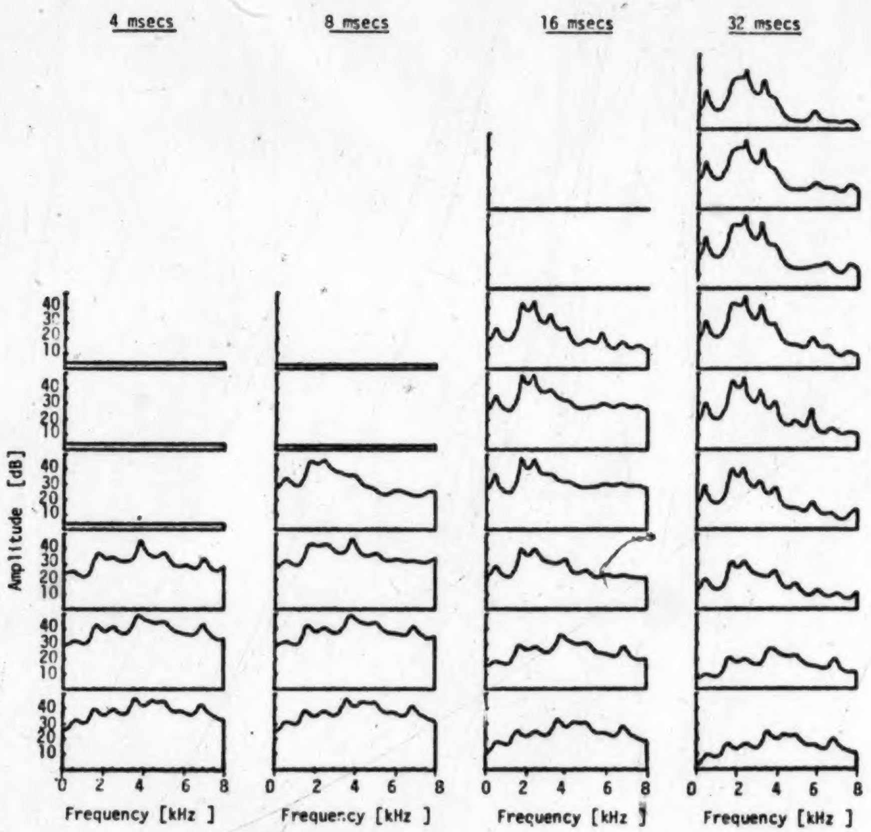
Syllable /bo/; Speaker PFA



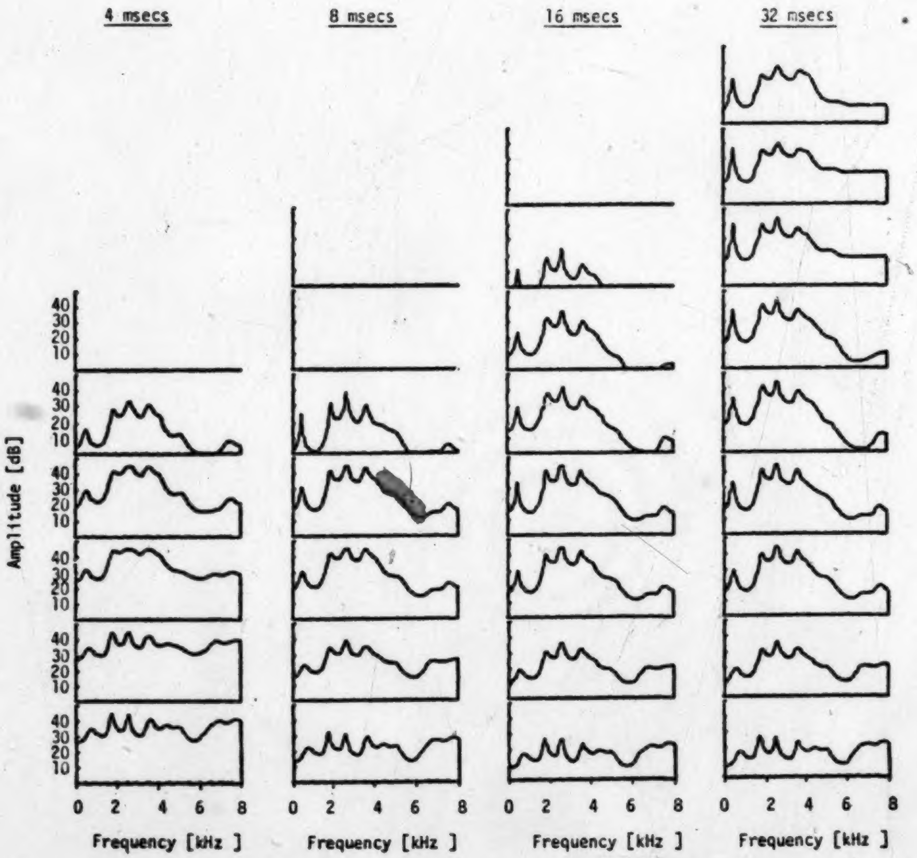
Syllable /bɛ/; Speaker PFA



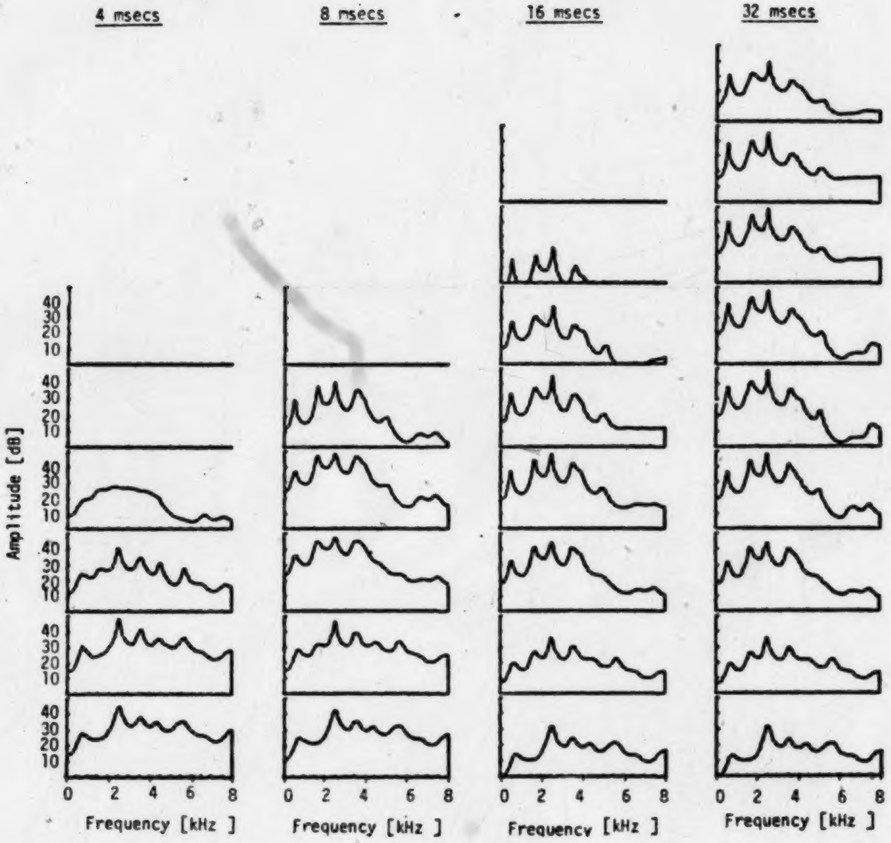
Syllable /be/; Speaker PFA



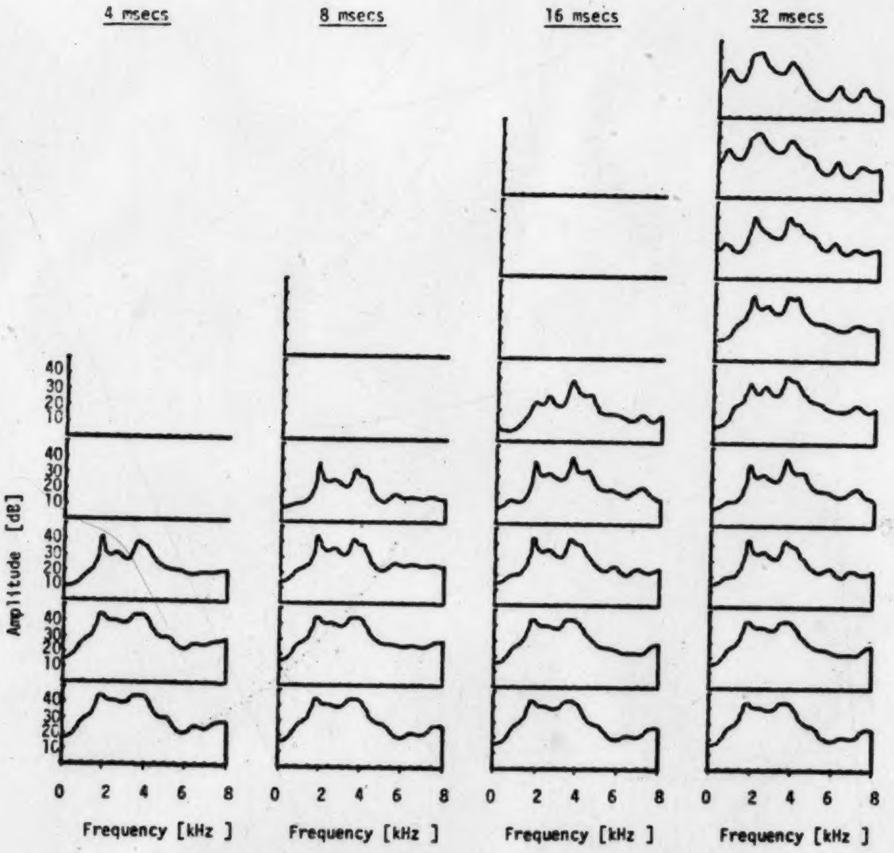
Syllable /bI/; Speaker PFA



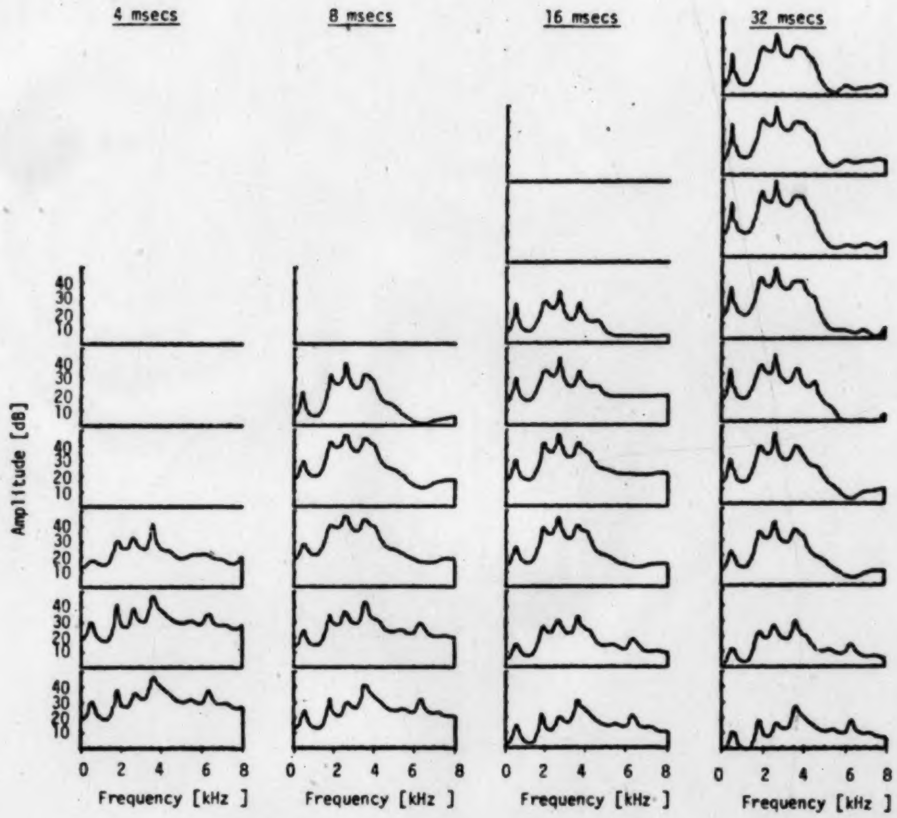
Syllable /bɛ/; Speaker RAH



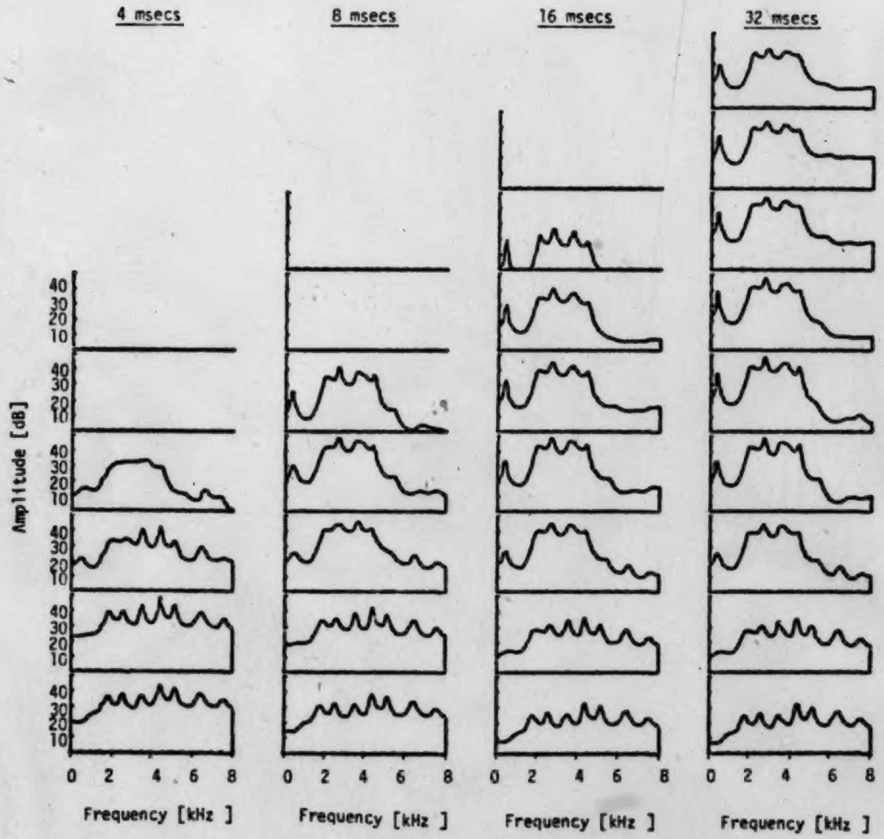
Syllable /bæ/; Speaker RAH



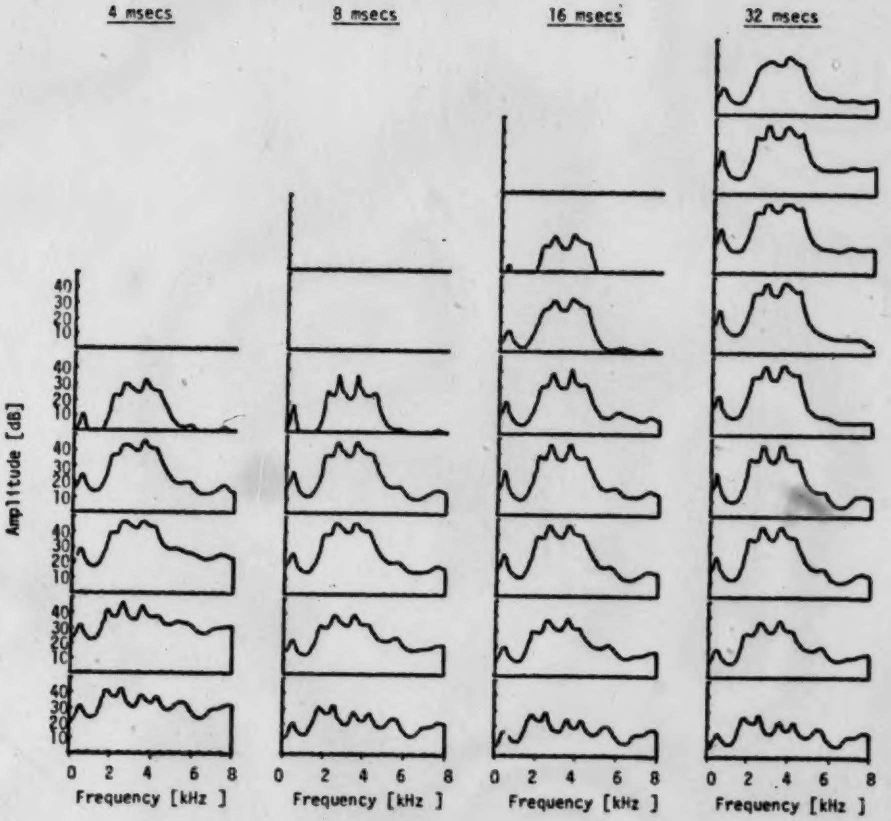
Syllable /gɝ/; Speaker RAH



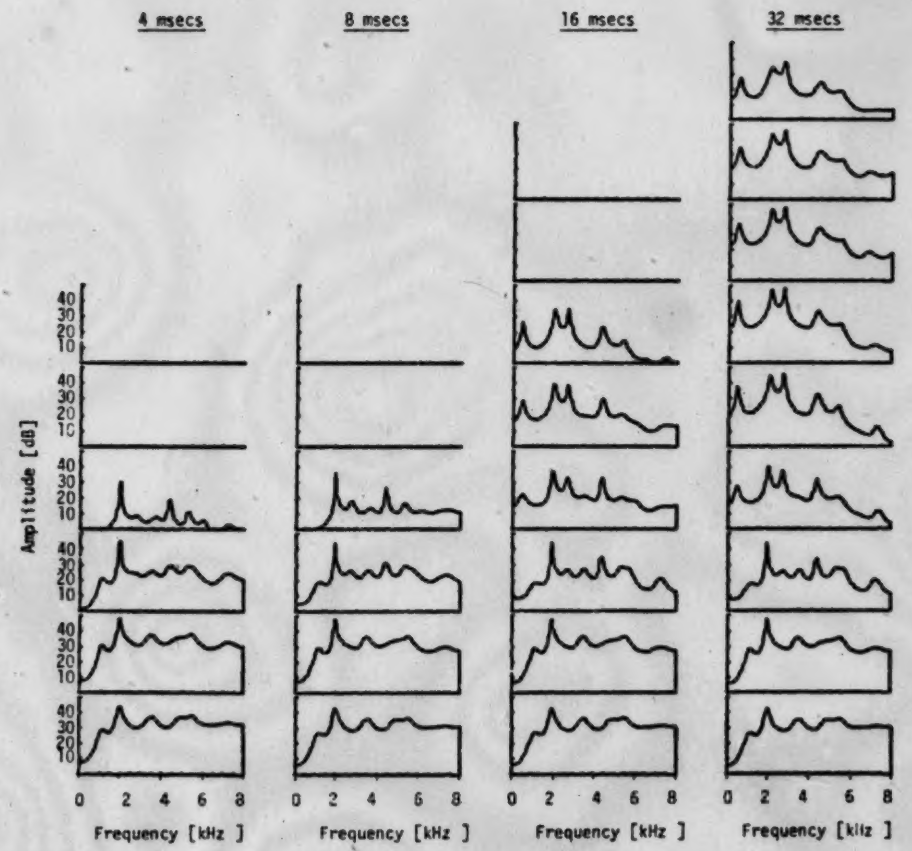
Syllable /be/; Speaker RAH



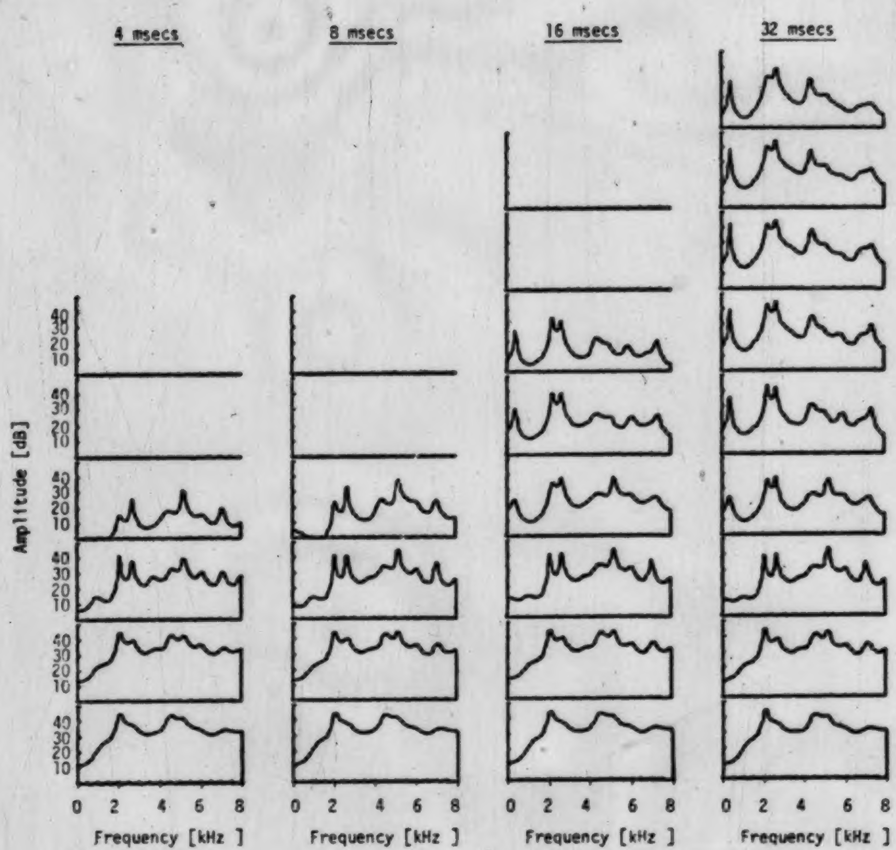
Syllable /bI/; Speaker RAH



Syllable /bi/; Speaker RAH

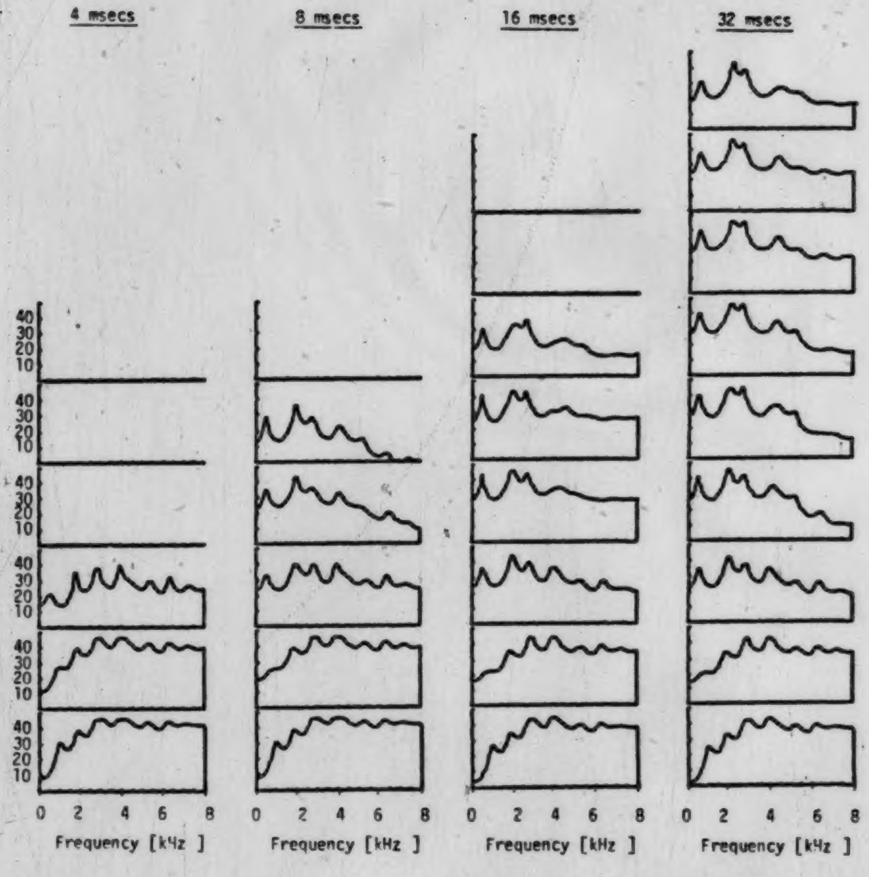


Syllable /dæ/; Speaker TMD

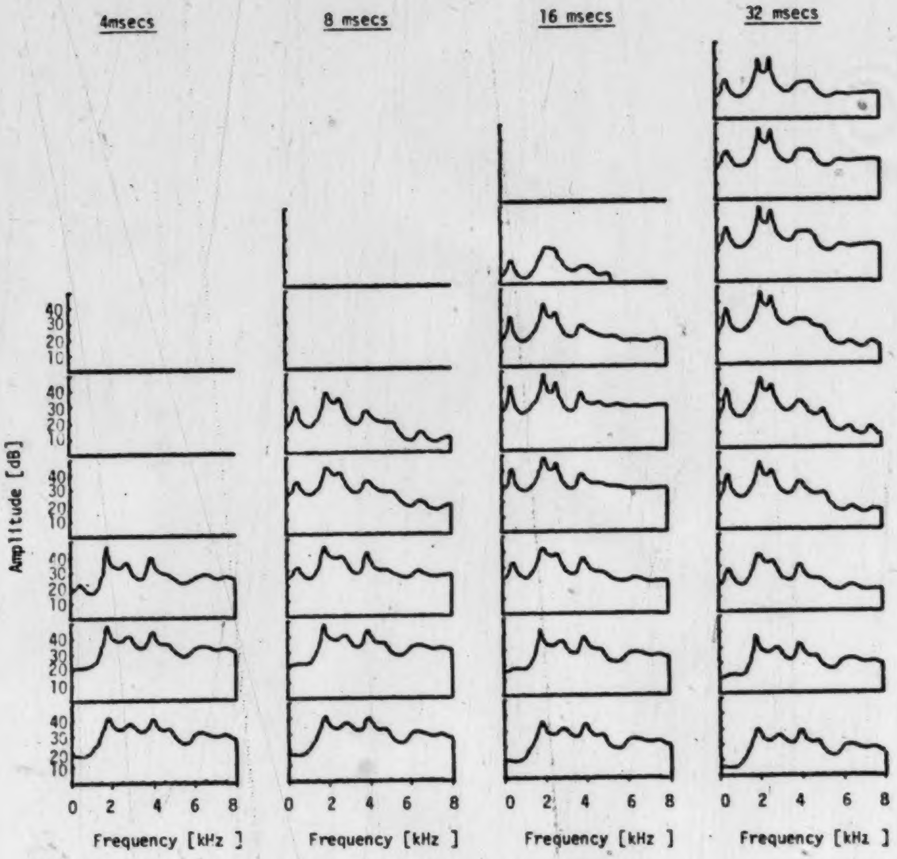


Syllable /dI/; Speaker TMD

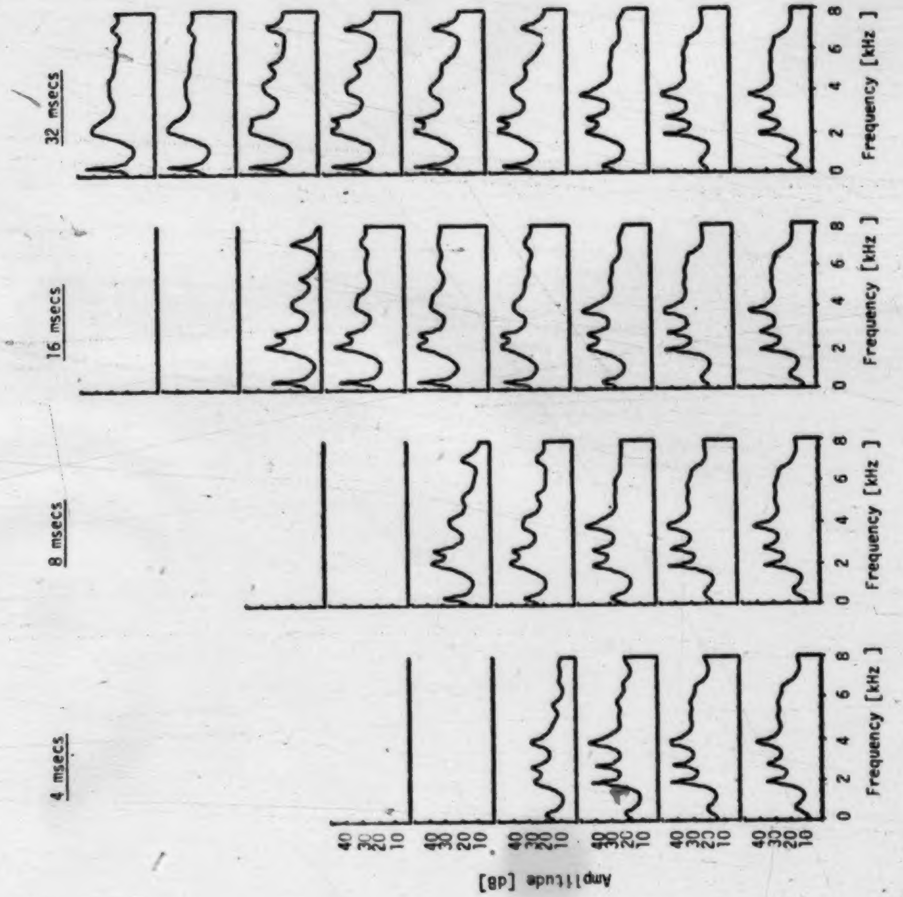
5



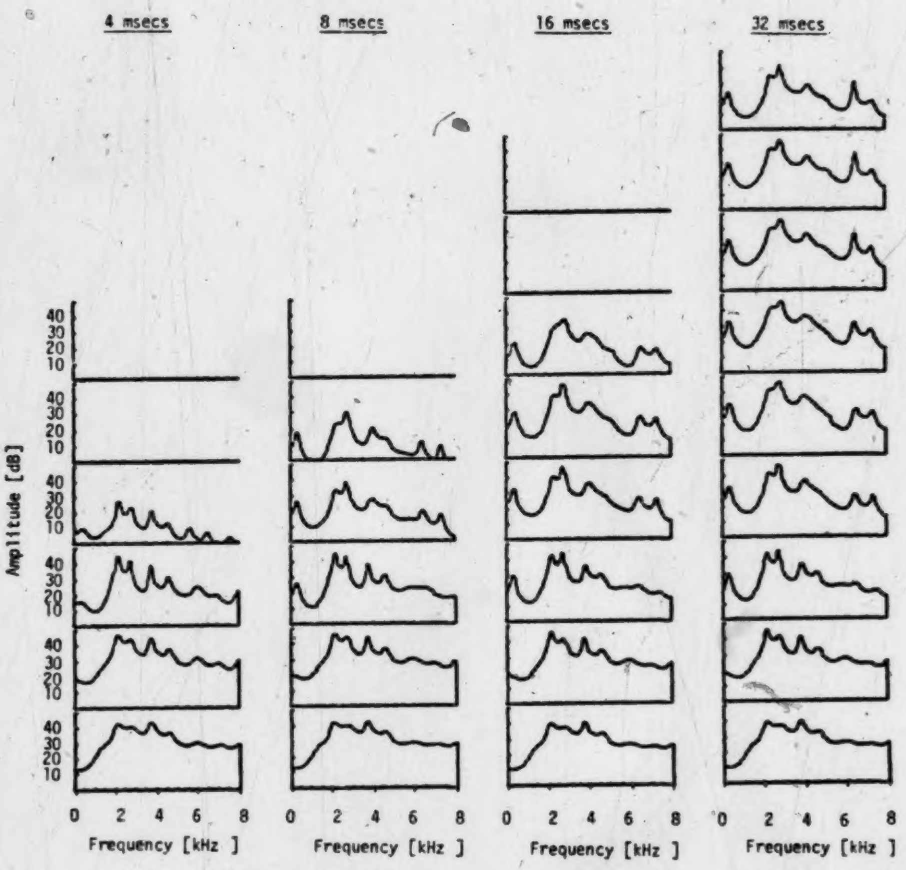
Syllable /bɛ/; Speaker TMD



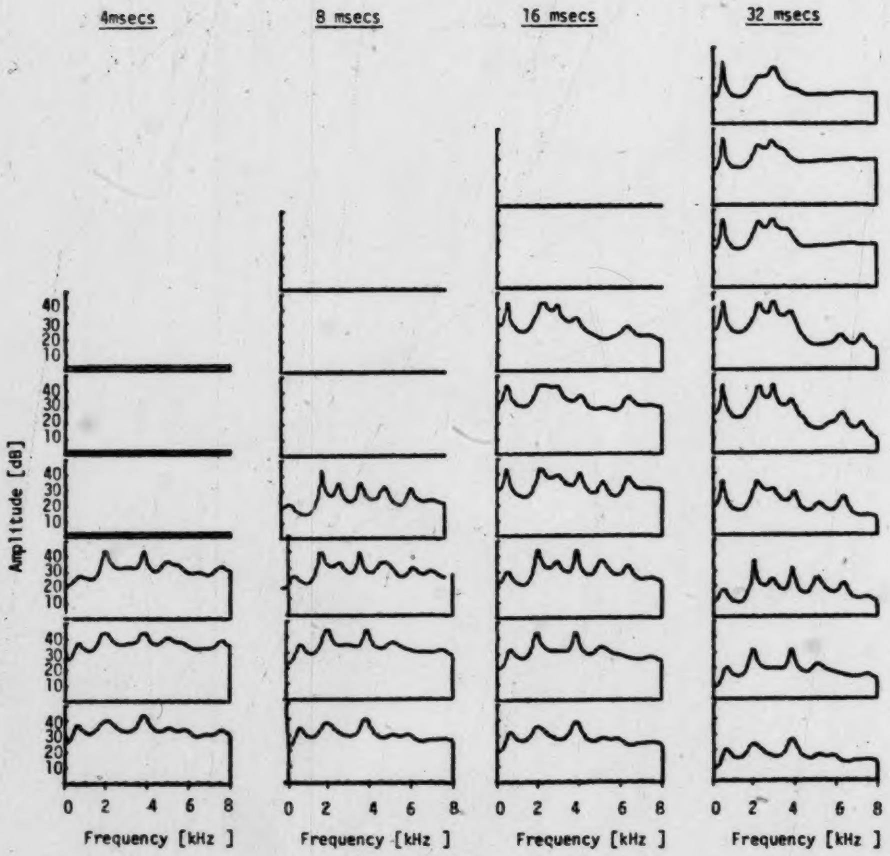
Syllable /be/; Speaker TMD



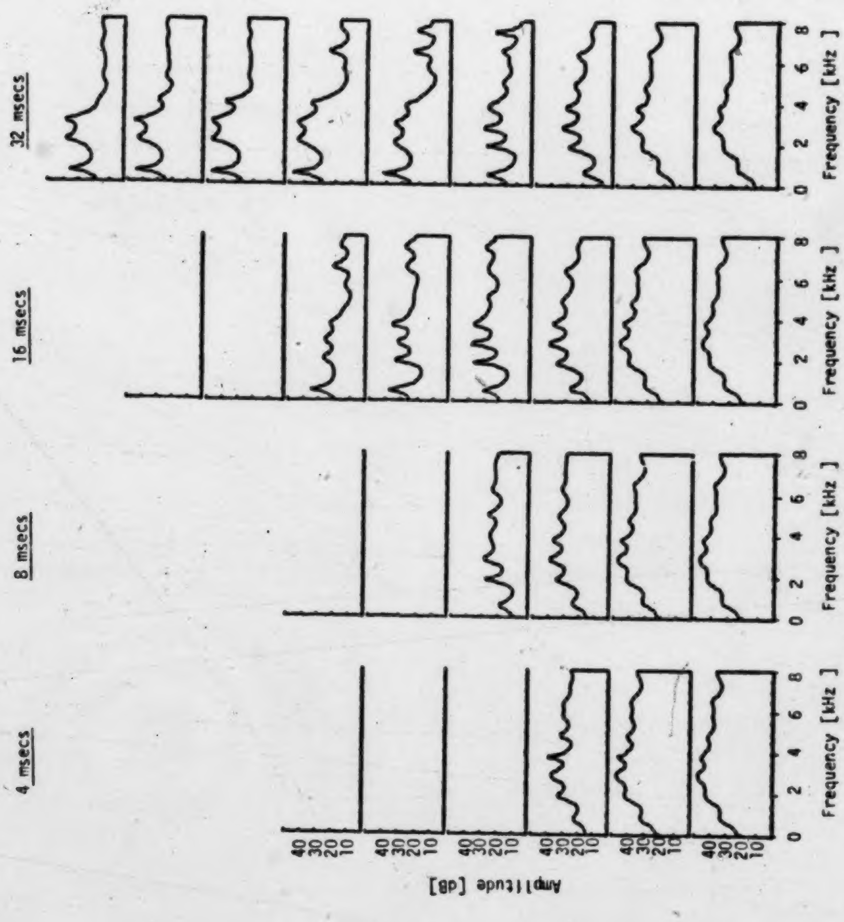
Syllable /bI/; Speaker TMD



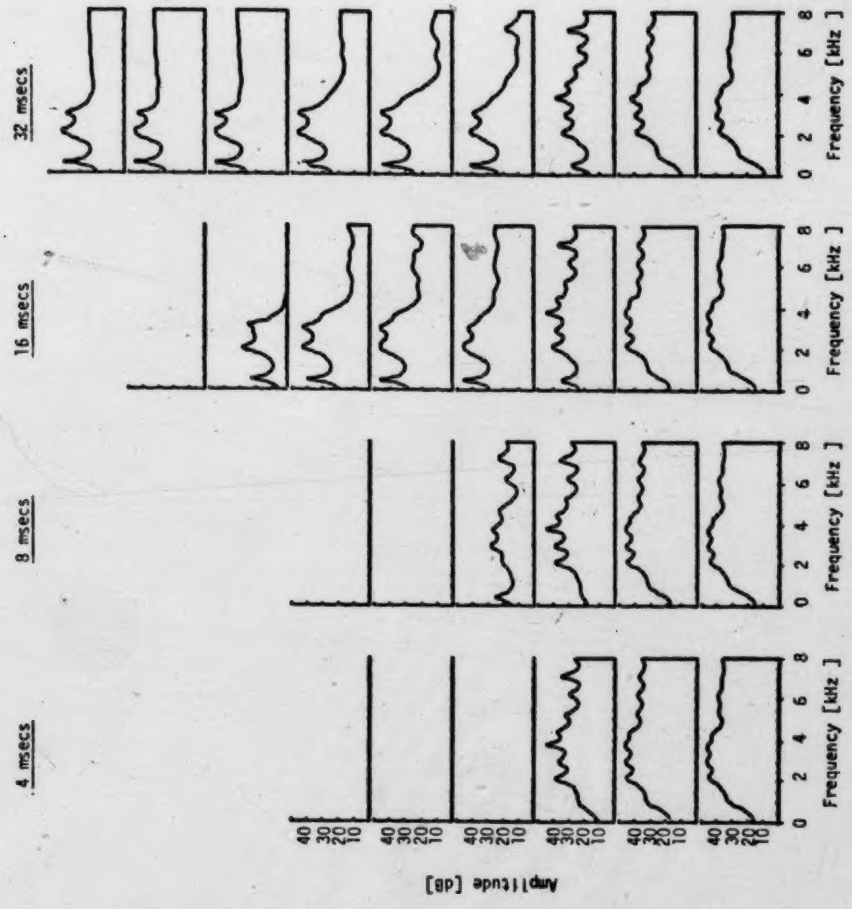
Syllable /bi/; Speaker TMD



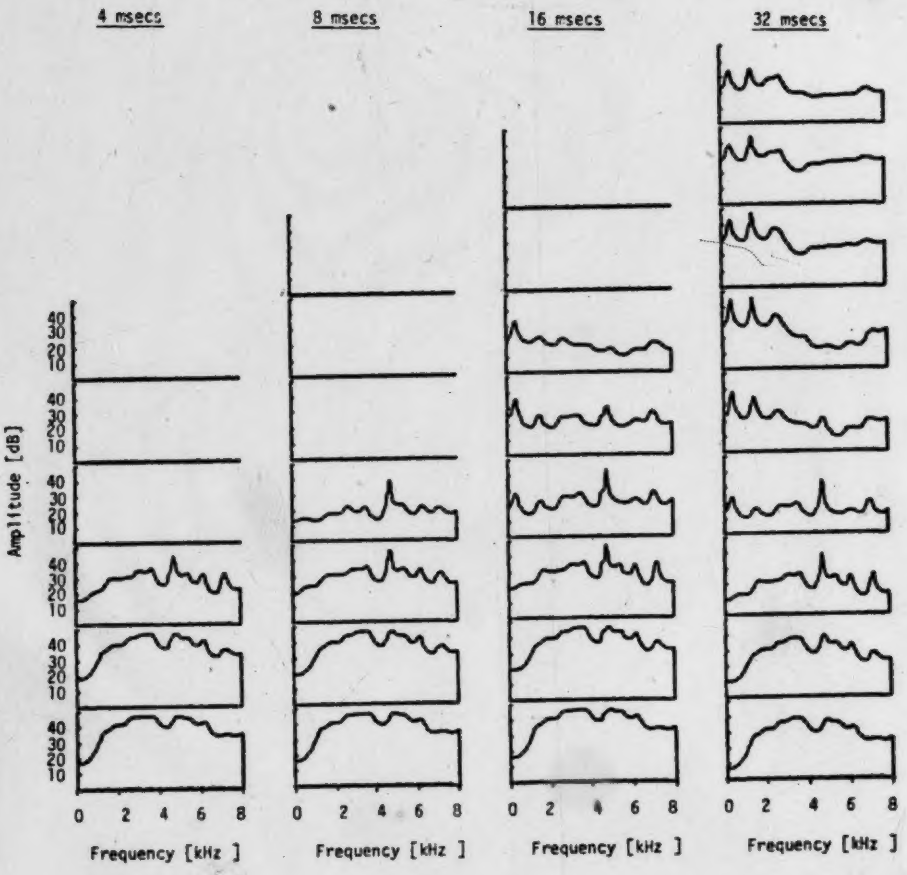
Syllable /be/; Speaker MLD



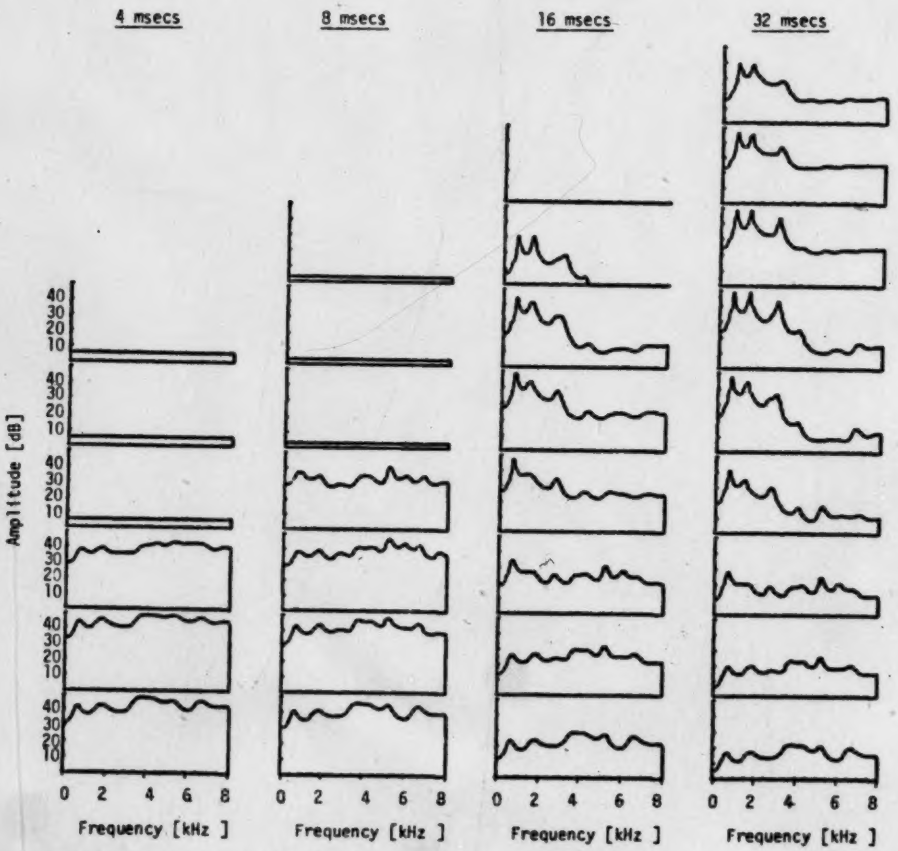
Syllable /bɛ/; Speaker MLD



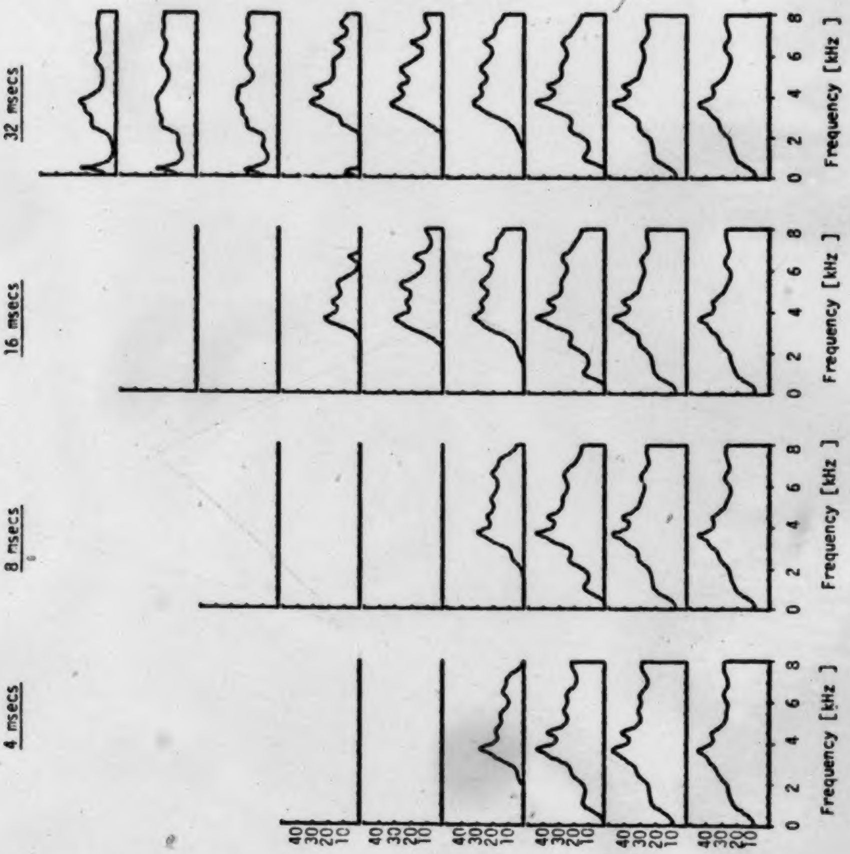
Syllable /dæ/; Speaker MLD



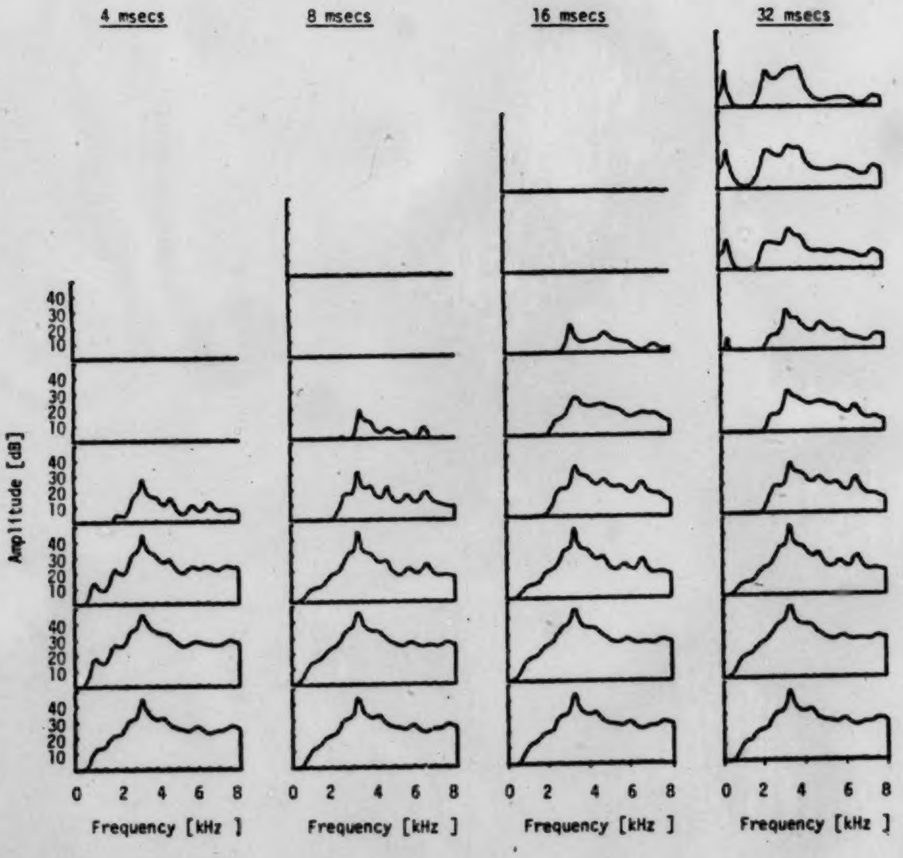
Syllable /bu/;  Speaker MLD



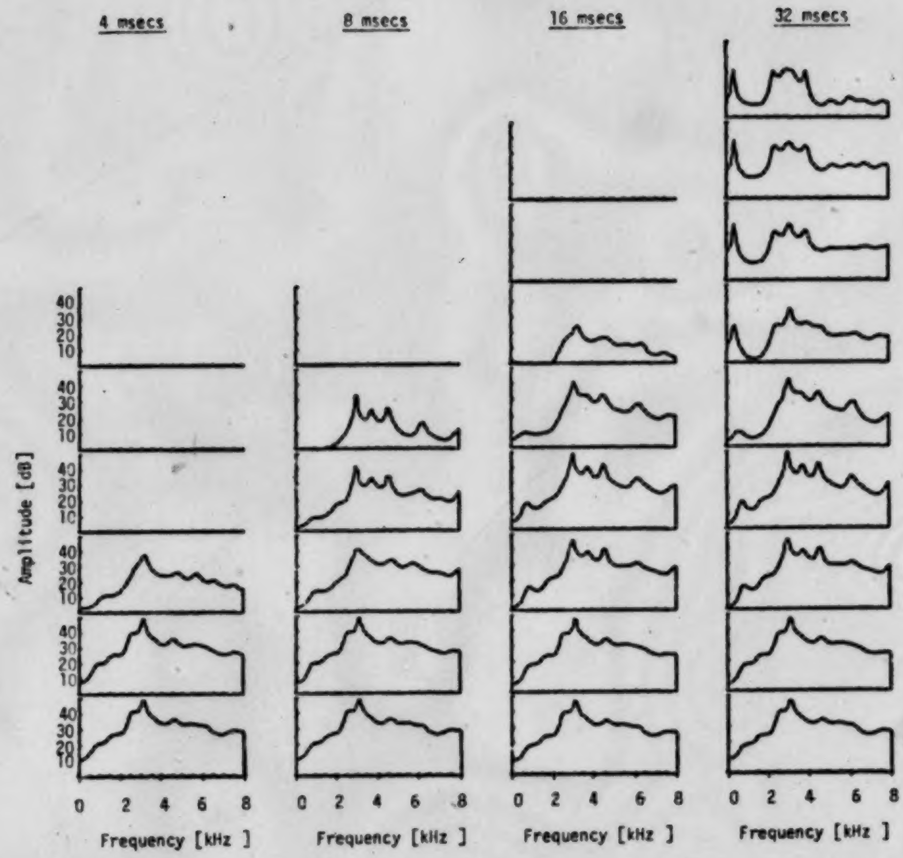
Syllable /bʌ/; Speaker MLD



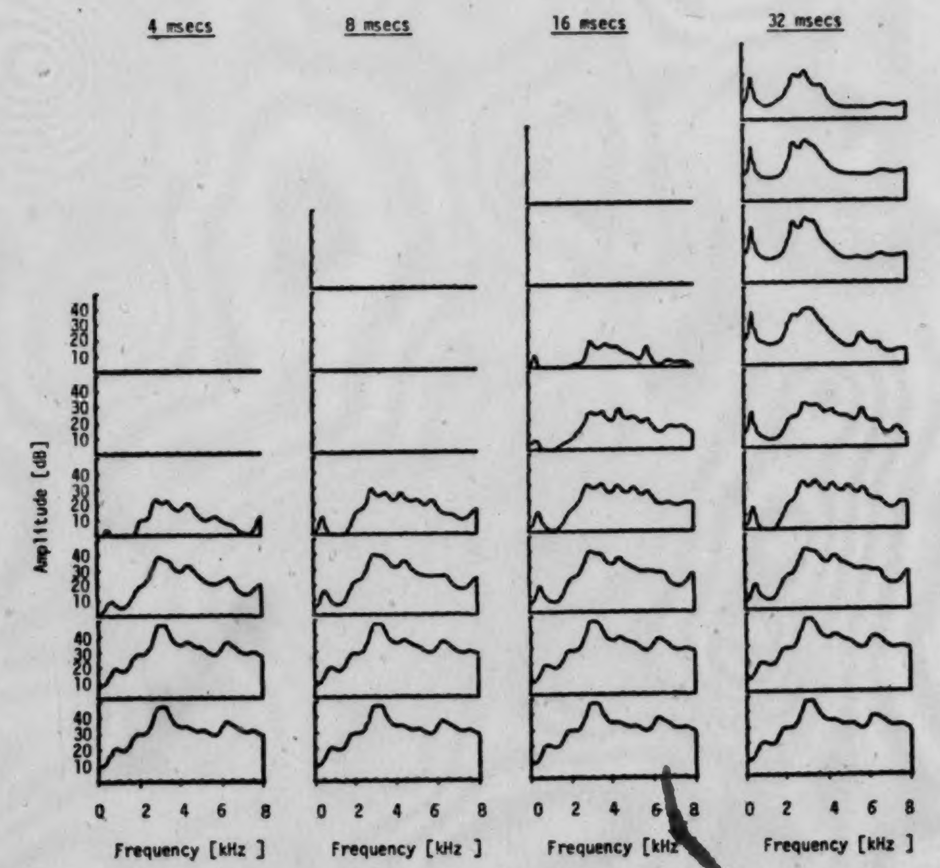
Syllable /gi/; Speaker MLD



Syllable /gI/; Speaker MLD



Syllable /ge/; Speaker MLD



Syllable /gɛ/; Speaker MLD